

THREE ESSAYS IN FINANCIAL ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by Alan Paul Kwan

May 2017

Copyright © 2017 Alan Paul Kwan

ALL RIGHTS RESERVED

THREE ESSAYS IN FINANCIAL ECONOMICS

Alan Paul Kwan, PhD

Cornell University 2017

This dissertation explores three different perspectives on frictions that impact the functioning of financial markets. My first and third essay explore information economics in financial markets. My second essay studies the role of regulatory scope and how financial regulation should be implemented.

In Chapter 1, “Does Social Media Cause Excess Comovement?”, I study social media’s potential to impact financial markets. When information is costly to produce, information intermediaries specialize in some stocks, creating flows of information and trading among such stocks. Trading by customers results in “excess”, seemingly non-fundamental comovement. Consistent with this theory, I find that co-mentioning of stocks explains increases in comovement. Three different empirical designs point toward a causal interpretation.

In Chapter 2 (joint with Chicago Booth PhD students Ben Charoenwong and Tarik Umar), “Who Should Regulate Investment Advisors?”, we study whether national or local regulators best deter investment adviser misconduct. Dodd-Frank provides us a laboratory to observe a large re-jurisdiction event in which state regulators below an arbitrary threshold were delegated to state regulation. Consistent with weakened regulation, customer complaint rates increase. The complaints represent more severe, not more frivolous reporting. Finally, they precipitate for firms and adviser representatives it might be assumed under the weakest oversight, such as those further from regulators.

In Chapter 3 (joint with Gaurav Kankanhalli and Kenneth Merkley), we study the disclosure paradox of Arrow (1962) – innovation requires secrecy, while financing or otherwise assessing economic value of innovation requires disclosure. Using a novel licensing database setting, in which many agreements are for a decade shrouded in secrecy, we compare the innovation paths of firms with originally redacted, but later unveiled, licensing activity versus those that don’t.

We find that as expected, firms that suppress information innovate more. However, the market, recognizing the problem of disclosure, appears to view redaction as a signal of quality, rewarding such firms with higher outside ownership and stock liquidity. We develop a signaling model to rationalize these viscerally puzzling results. Consistent with theoretical predictions, we find factors that reduce or increase adverse selection explain the differential effect on market outcomes of the redacting firms.

BIOGRAPHICAL SKETCH

Alan Paul Kwan was born in Los Angeles, CA to Paul and Cecilia Kwan, immigrants from Hong Kong. He has two older sisters, Agnes and Belinda Kwan. Alan grew up in a low-income neighborhood of Highland Park, CA. At age 14, his family moved to Alhambra, CA and Alan attended Alhambra High School. He was an active leader in several clubs and also achieved the rank of Eagle Scout. He graduated in 2005.

For college, Alan pursued his then dream of living on the East Coast and attended Dartmouth College. He participated heavily in clubs, volunteering and a variety of pre-professional activities. After three internships at the Center for Asian Americans United in Self-Empowerment, Bridge-water Associates, and Microsoft, he decided to pursue the first stage of his career in quantitative trading. In 2009, graduated with a degree in Economics modified with Mathematics.

After graduation, he briefly joined a startup consultancy, Edgeworth Economics. He then moved to the DC metropolitan area, where he worked as a quantitative analyst for a trading firm called DC Energy. Three years later, he joined Cornell University's doctoral program in finance, a half-decade experience which culminates in this dissertation.

This dissertation is dedicated to my parents, Paul Si Lung and Cecilia Wai Wan Kwan.

ACKNOWLEDGMENTS

The list of people who have contributed to my life during my PhD is too large to do justice here. However, I could not have, and would not have wanted to, completed my PhD without the motley assemblage of people that comprised my support network through this program.

First, I would like to thank my family. Reflecting for a moment, it is obvious what role my family has had in shaping my aspirations. My father's dabbling in computer science, law, accounting and personal finance are seemingly too coincidental to not be the source my own same interests. Meanwhile, my mother's work ethic in pursuing her graduate education while balancing a family and a career provided me a role model. She has told me she would pursue a doctorate if she were younger. Following the template for many children of diaspora, I have the opportunity my parents sacrificed coming to the United States.

Second, I am immeasurably grateful to Andrew Karolyi, my dissertation chair, who took a proactive interest in guiding me since my first year of graduate school. He emanates wisdom with every conversation and has perhaps the largest single-handed impact on how I conduct my research and professional self. I also thank every member of my committee, Byoung-Hyoun Hwang, Gideon Saar, and Scott Yonker. Their unwavering support, mentoring and warm manner has helped make this feel like a personal journey rather than a vocation. Also, although it is unlikely to be a unilateral decision, I associate Gideon with taking a bet on me admitting me to the program. I am also grateful to other faculty who've taken an interest in supporting PhD students intersecting with an impact on me: Matthew Baron, Elizabeth Berger, David Ng, Murillo Campello, Kenneth "NPH" Merkley, Roni Michaely, and Hyunseob Kim.

Last, but not least, I would like to thank a particular subset of my friends in alphabetical order: Bryce Little, Ben Matthies, Christine Chang, Christopher Chen, Cindy Lu, Da Woon Kim, Gaurav Kankanhalli, Jared Ye, Stephanie Lo, Yen Vu, Yolanda Lin and Yukun Liu. I did not name only people at Cornell, or in my field. More than serving as sounding board or muses for intellectual

inspiration, my friends have served to keep me even-keeled. They also tolerate the fact I don't visit or hang out as often as I plan to soon.

For Chapter 1, I would like to thank Andrew Karolyi for guiding me and keeping me on track. Every committee member of mine was exceedingly helpful. Gideon Saar gave me crucial methodological advice right before the job market. Byoung is a resident social media expert, and Scott encouraged me to aspire with this chapter. Both of these committee members are in their own ways some of the nicest people I know, and have helped me so much through the process. They all receive my highest recommendation to future students. I also would like to thank faculty at the Ohio State University and the University of Notre Dame, in particular, for helpful comments on how to pitch the paper. I also thank everyone from the interview trail who I met with and gave me their time listening me to explain my work.

For Chapter 2, I am grateful to Scott Yonker for unknowingly seeding me the idea. I am also grateful for feedback from various faculty at Cornell and the University of Chicago.

For Chapter 3, I would like to thank Murillo Campello, and David Brown of the University of Florida, for providing helpful suggestions that resulted in direct and meaningful changes to the paper. Ednaldo Silva of RoyaltyStat is also someone who earns my eternal gratitude for sharing me business knowledge and data that together comprise the opportunity of a lifetime.

I thank you all for making an imprint on my work, including those many I have not named. The past five years have flown by. It is said, however, that time flies when you're having fun!

Contents

| | |
|--|------------|
| Preface | ii |
| Biographical Sketch | iii |
| Acknowledgements | v |
| 1 Does Social Media Cause Excess Comovement? | 1 |
| 1.1 Introduction | 1 |
| 1.2 Data | 8 |
| 1.2.1 About StockTwits | 8 |
| 1.2.2 Other data | 11 |
| 1.2.3 Calculating correlations | 12 |
| 1.2.4 Social media summary statistics | 13 |
| 1.3 Methodology | 15 |
| 1.3.1 Empirical design | 15 |
| 1.3.2 Sample selection | 21 |
| 1.3.3 Sample summary statistics | 23 |
| 1.4 Baseline correlations | 26 |
| 1.4.1 Baseline panel regression | 26 |
| 1.4.2 Robustness - is social media activity a substitute for news? | 28 |
| 1.4.3 Robustness - timing of posts | 31 |
| 1.5 Common follower networks as an instrumental variable | 33 |
| 1.5.1 Design | 33 |
| 1.5.2 Social graph data | 34 |
| 1.5.3 Results | 35 |
| 1.5.4 Endogeneity concerns regarding social connections | 39 |
| 1.6 Instrumental variable based on StockTwits banner | 40 |
| 1.6.1 Exclusion restriction | 40 |
| 1.6.2 Data, sample period, and design | 42 |
| 1.6.3 Results | 43 |
| 1.7 Other tests | 46 |
| 1.7.1 Followership | 46 |
| 1.7.2 Stock characteristics | 48 |
| 1.7.3 Investor-reported sentiment | 54 |
| 1.7.4 Other types of comovement | 55 |
| 1.8 Conclusion | 58 |

| | | |
|----------|--|------------|
| 2 | Who Should Regulate Investment Advisors? | 60 |
| 2.1 | Introduction | 60 |
| 2.2 | Oversight of Investment Advisers and Broker-Dealers | 64 |
| 2.2.1 | History of Regulatory Jurisdiction | 64 |
| 2.3 | Data & Methodology | 70 |
| 2.3.1 | Data collection | 70 |
| 2.3.2 | Sample Construction | 71 |
| 2.3.3 | Methodology | 72 |
| 2.3.4 | Identifying Treatment and Control Groups | 73 |
| 2.3.5 | Summary Statistics | 74 |
| 2.4 | Results | 78 |
| 2.4.1 | Complaint Incidence | 78 |
| 2.4.2 | Types of Complaints | 84 |
| 2.4.3 | Recidivism | 85 |
| 2.5 | Misconduct or Reporting? | 86 |
| 2.5.1 | Serious Complaints | 86 |
| 2.5.2 | Under-staffed State Regulators | 91 |
| 2.5.3 | Dilution of Regulatory Resources Impacting Incumbent Advisers | 94 |
| 2.5.4 | Distance from Regulators Impacting Oversight | 94 |
| 2.5.5 | Client Composition and Treatment Response | 98 |
| 2.6 | Conclusion | 101 |
| | | |
| 3 | Speech is Silver but Silence Is Golden: Information Suppression and the Promotion of Innovation | 103 |
| 3.1 | Introduction | 103 |
| 3.2 | Institutional Context and Hypothesis Development | 109 |
| 3.2.1 | Institutional Context | 109 |
| 3.3 | Model and Hypothesis Development | 111 |
| 3.4 | Data | 116 |
| 3.4.1 | Intellectual Property Licensing Data | 116 |
| 3.4.2 | Other Data | 120 |
| 3.4.3 | Summary Statistics | 122 |
| 3.5 | Methodology | 134 |
| 3.6 | Consequences of Strategic Redaction | 138 |
| 3.6.1 | Redaction and Capital Market Outcomes | 138 |
| 3.6.2 | Redaction and Future Innovation | 144 |
| 3.6.3 | Cross-sectional Heterogeneity | 148 |
| 3.7 | Conclusion | 154 |
| | | |
| A | APPENDIX (CHAPTER 1) | 161 |
| | | |
| B | APPENDIX (CHAPTER 2) | 170 |
| B.1 | Redacting Customer Complaints | 171 |

| | |
|---|------------|
| C APPENDIX (CHAPTER 3) | 174 |
| C.1 Full Model Analysis | 174 |
| C.2 Sample Construction and Selection | 176 |
| C.3 Control Variable Definitions | 178 |
| C.4 Robustness Tests | 179 |

List of Tables

| | |
|---|-----|
| 1.1 StockTwits and Message Volumes | 16 |
| 1.2 Cross-sectional Distribution of Posters' Characteristics | 22 |
| 1.3 Sample summary statistics | 24 |
| 1.4 Baseline correlations | 27 |
| 1.5 Robustness - Co-postings during News Events | 29 |
| 1.6 Instrumental variable based on on common follower networks | 36 |
| 1.7 Instrumental variable regressions at the weekly horizon using the StockTwits banner | 44 |
| 1.8 Do posters with more followers have greater influence? | 47 |
| 1.9 Is there a larger effect on high arbitrage cost stocks? | 49 |
| 1.10 Investor style classes | 50 |
| 1.11 Sentiment | 53 |
| 1.12 Other types of comovement | 56 |
| 2.1 Observation counts | 75 |
| 2.2 Firm Summary Statistics | 76 |
| 2.3 Summary Statistics around Dodd-Frank | 77 |
| 2.4 Baseline Results | 80 |
| 2.5 Robustness: Alternative Comparison Groups | 81 |
| 2.6 Different AUM cutoffs | 83 |
| 2.7 Customer Complaint Decomposition | 87 |
| 2.8 The Effect of Regulatory Jurisdiction on Recidivism | 88 |
| 2.9 Alleged Damages Analysis | 90 |
| 2.10 Complaint Noise-to-Signal Ratio | 90 |
| 2.11 Staffing of the Investment Adviser Regulatory Office | 92 |
| 2.12 Treatment Effect on Existing State-Registered Firms | 95 |
| 2.13 Distance to Regulator | 96 |
| 2.14 Client Composition | 97 |
| 2.15 Client Composition | 99 |
| 2.16 Client Composition at the Branch Level | 100 |
| 3.1 Data Fields Extracted from RoyaltyStat Data | 117 |
| 3.2 Sample Selection Procedure | 120 |
| 3.3 Contract-level Summary Statistics | 123 |
| 3.4 Firm-Level Summary Statistics | 127 |
| 3.5 Benchmark Model of Determinants of Redaction | 131 |
| 3.8 Effect of Competition on Likelihood of Redaction | 133 |
| 3.9 Effect of Agency Problems on Likelihood of Redaction | 134 |
| 3.10 Percentage Improvement in Balance as a Result of Matching | 137 |
| 3.11 Effect of Redaction on Future Stock Liquidity | 139 |

| | | |
|------|---|-----|
| 3.12 | Effect of Redaction on Future Equity Issuance and Institutional Ownership | 141 |
| 3.13 | Effect of Redaction on Future Patenting | 145 |
| 3.14 | Effect of Redaction on Future Intangible Capital Accumulation and R&D Expense | 149 |
| 3.15 | Effect of Redaction for Small and Young Firms | 150 |
| 3.16 | Effect of Redaction Conditional on Past Redaction | 151 |
| 3.17 | Effect of Redaction Conditional on Equity Dependence | 153 |
| A.1 | Sample Selection Process | 161 |
| A.2 | Variable Definitions | 162 |
| A.3 | Average Annual Stock Characteristics | 163 |
| A.4 | Most Popular Stocks in 2015 | 163 |
| A.5 | Additional summary statistics | 164 |
| A.6 | User-level coverage statistics | 165 |
| A.7 | Timing decomposition | 166 |
| B.1 | State Security Regulators | 170 |
| B.2 | State Security Regulators (continued) | 171 |
| C.1 | Effect of Redaction on Future Liquidity in Event-Time | 179 |
| C.2 | Robustness with SIC2-by-Year Fixed-Effects | 180 |

List of Figures

| | | |
|-----|---|-----|
| 1.1 | Stocktwits user interface | 9 |
| 1.2 | Post Distribution Throughout Trading Day | 15 |
| 1.3 | Sample Common Follower Calculation For Identification Strategy | 32 |
| 2.1 | Event Timeline | 68 |
| 2.2 | Annual ADV Deregistration Filings | 68 |
| 2.3 | Filing a Customer Complaint | 69 |
| 2.4 | Complaint Example & Summary Statistics | 70 |
| 2.5 | Parallel Trends | 82 |
| 2.6 | State Regulator Budgets | 93 |
| 3.1 | Sample Redacted Licensing Agreement | 112 |
| 3.2 | Distribution of Contracts by Technology Industry | 125 |
| 3.3 | Distribution of Redacted vs. Un-redacted Contracts Over Time | 126 |
| 3.4 | Relationship Between Redaction Rate and Proprietary Costs of Disclosure | 129 |
| A.1 | StockTwits integration with Yahoo! Finance and Marketwatch | 167 |
| A.2 | Follower network over time | 169 |

Chapter 1

Does Social Media Cause Excess Comovement?

1.1 Introduction

Standard asset pricing theory posits that asset prices comove because they have common exposures to state variables or other economic factors. Yet there is a growing literature that suggests return comovement is generated by forces unrelated to pure fundamentals. Correlated trading activity by institutional investors, for instance, can drive comovement as common ownership of stocks increases.¹ Another driver of comovement appears to be the selective, specialized production of information by information intermediaries, such as equity analysts and equity underwriters, who possess costly-to-obtain private information about only special subsets of the equity universe.² In contrast, others argue that psychological biases pervasive across investors can imbue certain stocks with special appeal, or transitory events make subsets of the stock universe salient to unsophisticated investors, resulting in excess comovement due to investor demand shocks.³

I propose that social media could generate excess comovement. Social media is an interesting setting because of its widespread adoption in the last decade. While internet communications platforms such as message boards have been around and studied in finance for almost two decades, modern social media is fundamentally different - it is now a pervasive, entrenched societal institution.⁴ In September 2015, over a billion people worldwide used Facebook in a *single day*, and today, perhaps one-third of the world is on a social media network of some kind.⁵

The adoption of social media in the financial services industry is well underway. SocialMe-

¹Barberis et al. (2005) and Greenwood (2008) study index inclusions. Anton and Polk (2014) study institutional ownership, a finding confirmed by Ben-David et al. (2016). Bartram et al. (2015) study an international context. Basak and Pavlova (2013) rationalize comovement through index-tracking motives of institutional investors.

²Israelsen (2014) and Hameed et al. (2015) discuss equity analysts. Grullon, Underwood and Weston (2014) points to stock underwriters, and Scherbina and Shluse (2015) points to news media, as drivers of comovement.

³The landmark paper in this is Barberis et al. (2005). Green and Hwang (2009) study nominal price categories. Kumar and Lee (2006) show retail order imbalances cluster in time. Kumar et al. (2013) and Kumar et al. (2013) study co-movement among lottery stocks and within geographies.

⁴A non-exhaustive list includes studies such as Antweiler and Frank (2004) and Tumarkin and Whitelaw (2001).

⁵WeAreSocial.Net reports that in August 2014, worldwide users of social media surpassed 2 billion across major social media platforms, such as Facebook, QQ, WeChat, etc.

diaAnalytics, a data vendor devoted to social media in finance applications, uses platforms like StockTwits and Twitter to generate trading signals for clients. In addition, in 2014, BNY Mellon issued a survey to companies, finding that 27% of companies around the world use social media actively to engage investors, complementing findings from Jung et al. (2015), who find 63% of the S&P 1500 use Twitter to supplement earnings releases. With large quantities of content being generated on social media and the essentially free cost of idea dissemination, the implications of social media for financial markets need to be understood.

Social media's rapid rise makes it not only an interesting setting, but also an *ideal* setting to test theories of excess comovement. Given the vast amounts of content produced by social media, it is plausible to interpret social media as a mediator of information. Veldkamp (2006) theorizes that when information is costly to produce, certain subsets of information - especially relatively inexpensive information - will be oversubscribed, generating excess comovement. Social media is in this sense "cheap" - social media platforms socialize the cost of procuring information and create a public good available to any interested internet user. Indeed, increasingly, surveys by think-tanks and industry groups suggest that social media is increasingly the preferred starting point for individuals to obtain or find news. Meanwhile, traditional news media continues to see rapid economic decline, with social media sometimes credited with either providing a substitute, or redirecting traffic that would traditionally go to news media.⁶

To measure social media activity, I use proprietary data from StockTwits. Since its launch in 2009, StockTwits has been a pioneer in the financial industry, as the first platform to invent the "\$cashtag" convention widely used by other social media sites, and the first platform to offer itself as a platform for company earnings releases. Today, StockTwits sits among the top as a market leader in financial social media. Through its large partnership list, throughout time encompassing prominent media entities such as Yahoo! Finance, Marketwatch, Bloomberg and Reuters, Stock-

⁶A report by the Brunswick Group (2012) suggests that 52% of surveyed respondents acquired information from investments from blogs and a quarter use social media to help make an investment decision. The Pew Research Institute, more generally, comments that one-third of individuals under 30 use social media as their primary source of news. Other resources such as Adweek.com echo similar numbers, such as "since 2009, 57% of adults get their news through Facebook and Twitter". Please see this link.

Twits in 2012 claimed to reach 100 million people worldwide, supplementing its own user base of 250,000 users in 2012, which ballooned to 1 million in 2015 and 1.6 million monthly active users in 2016. AlexaRank suggests in 2015 that StockTwits is one of the 20 most prominent investing-related websites in the world.⁷

Using this setting, I study the relation between social media activity and stock return correlation. For this study, StockTwits provided me with the full survivorship-bias free set of postings on its website, consisting of 37 million postings referencing a stock ticker through June 2016. The data contain the time, subject, investor-reported sentiment and data about the users' identities. I supplement this data with data I gathered from webcrawling the "social graph" of every poster on the site, as well as from live screen capturing of the site over 10 months.

I first create a stock-pair-time panel, consisting of a subset of the stock pair universe. I then run two parallel sets of tests. In my primary test, I study monthly correlations calculated from Fama and French (2016) five-factor-adjusted daily returns. In a complementary set of tests, I shorten the horizon to a calendar week and then study intraday, half-hour, market-adjusted return correlations. While the former measures comovement at an investor-relevant horizon, the latter befits the frenetic pace of social media.

In my baseline correlations, I show that at the monthly and weekly horizons, there is a strong relation between social media activity and excess return correlation. In a conservative within-pair, within-time framework, factor-adjusted correlation increases by 104 basis points at the monthly level, when social media activity goes from zero to 3 stock pair posts. At a 30 minute, intraday level, the effect is 34.6 basis points. This is economically large in standalone terms, still when considering the 25% level in monthly raw return correlation, and more so when considering the 1.5% average excess correlation. The effect is not explained a basic battery of controls, nor is it simply explained by news announcements. In fact, the effect of social media activity on return

⁷Retrieved July 7, 2015. In September, 2016 SimilarWeb suggests StockTwits' worldwide category rank is 41, a source which the StockTwits head of development suggested was probably the most reliable. For reference, the category leader is Yahoo! Finance, a StockTwits partner during my sample period. Offhand comparisons of the SimilarWeb numbers and numbers he revealed to me suggest either Alexa or SimilarWeb are likely underestimates of true traffic to this site.

correlation is in fact weaker during periods when exogenous news events happen to coincide (as captured by Capital IQ's Key Developments database), and barely budges after adding industry-pair-time fixed effects.

Still, the natural concern with this analysis is that events that drive correlations may also be observed by social media posters, in the manner of the Manski (1993) reflection problem. To argue causality, I present two causal experiments and a battery of supportive cross-sectional tests. The two causal experiments attempt to assign the pairs of stocks that posters co-mention. The experiments are based on the idea that users may be influenced by the way social media idiosyncratically presents a pair of stocks at the same time, increasing their salience.

In my first experiment, I use social networks, a generic feature of many social media platforms. The logic of the experiment is that while at any point in time, any post may be endogenous to an event outside of social media, the way the posts are aggregated by social media is potentially idiosyncratic. Specifically, I use the number of poster-pairs discussing *different* stocks who have *ex ante common followers*. I scrape the entire StockTwits website to collect and infer posters' historical followership lists. I compute only ex-ante common followers from at least six-months prior to a post, pre-empting any concerns over serially correlated news event or social media consumers' anticipation of future events.

Given two posters independently discuss two different stocks, that they share a follower in common *increases* the likelihood that some other user will co-mention these two stocks. Therefore, given two stocks of the same contemporaneous posting, the count of common followers serves as a valid instrument for co-posting. That is, controlling for the unconditional level of posting, the *instrumented* co-posting is attributable solely to the percolation of ideas through social media. In the first stage, I find that the relation between the number of co-views is highly positive. The instrumental variable achieves an F-statistic at the monthly level of 34.34 at the monthly level, and 54.22 at the weekly level. In the second stage, the effect is extremely robust at the monthly level; assuming, conservatively, that the only effect comes through these intransitive triads, a 25th percentile to 75th percentile increase in social media activity corresponds to between 1% and 1.5%

at the monthly level. The effect reduces, but remains mostly reliable, at the weekly level to about 23.4 basis points (up to about 45), even despite intraday noise.

My second causal experiment uses a strategy that is specific to the StockTwits platform. I instrument the co-posting intensity on a pair of stocks at the weekly level using the StockTwits banner, a ticker at the top of the StockTwits web page that comprises one of its most prominent institutional features. Every 5 minutes, StockTwits sorts stocks by message volume, transformed by a function that takes into account the 24 hour message volume and stock-specific parameters. This process is not divulged to the public or even partners, but creates significant distortions: the most popular stock is Apple Inc. with 5% of message volume, but \$AAPL is on the banner during my sample only 0.5% of the time. Thus, I argue that subject to controls for message volume, stock volume, turnover, stock sentiment and many higher order terms of stock returns, the within-pair, within-week display of a stock on the board is quasi-random. I explore this instrumental variable in a weekly test, as I recorded a dataset of ten second snapshots of the banner over a 10 month period.

Under many conservative, parsimonious experimental designs, I show that the synchronous display of two stocks on the StockTwits banner is highly relevant for the within-pair level of co-postings on a stock-pair. Moreover, not only is being on the board relevant, but *where* two stocks are on the board *visually* is relevant: stocks ranked higher on the board (further left, rather than further right) garner more attention, and stocks spaced visually further apart receive fewer co-postings, all else equal. In the second stage, I provide moderate evidence that instrumented co-postings positively relate to intraday return correlation. In my most rigorous specification, I show that, even if being on the StockTwits banner is an endogenous event entering the second stage, simply being further left on the board is relevant for the intensity of co-postings, and that instrumented intensity positively relates to excess return correlation.

Third, and finally, to the extent social media *cause* common movements in market activity and outcomes, five logical corollaries arise, each of which I confirm. First, consistent with the notion of poster influence, I show that posters with larger followings, otherwise equal, are 2 to 3.1 times more

influential, depending on stock sorting characteristic. Second, decomposing stocks by type, I show that for those pairs who are “vulnerable” to retail investors in the sense of Llorente et al. (2002) or Baker and Wurgler (2006), the relationship between social media activity and return correlation is two to three times larger. In addition, I show that stocks that are similar by “style” class in the sense of Barberis and Shleifer (2003) also comove more when social media activity is high. I study five style classes industry, price, value, momentum, market beta and investment-to-assets, finding that return correlation responds to social media activity less when the two stocks are in different style classes, such that the extremes of some style classes, the effect is fully abated. Fourth, I interact the main effect with investor-reported sentiment, measured using StockTwits’ commonly used feature, “the cashtag”. Defined in three ways, disagreement in the direction of sentiment on the components of a stock pair leads to a reduction in the effect of posting on comovement - one-third at the monthly level, and almost negative at the weekly level (although statistically unreliable). While sentiment is not exogenous, this is consistent with the idea that social media consumers trade according to their reported sentiment, and therefore impact returns accordingly. Finally, I repeat the analysis briefly on comovement in trading activity, absolute returns and liquidity, and find evidence suggesting even larger increases in these other market quantities.

This chapter is at the intersection of the burgeoning literature on the role of the news media, and social media, in financial markets, and the budding field Hirshleifer (2014) terms “social finance,” which aims to study how “ideas spread and evolve, and how social processes affect financial outcomes”. In this chapter, I present a setting in which I can directly track the microstructure of social transactions, which many prior studies can observe only indirectly, *and* demonstrate that these social transactions impact financial markets.⁸ Meanwhile, most research on social media platforms in finance has traditionally studied social media as an information source. The vast majority of

⁸Please see Hong et al. (2004), K Pool et al. (2014), or Mitton et al. (2015), who use proximity as a proxy for social interactions to explain correlated behavior among actors from certain geographies. Simon and Heimer (2012) and Heimer (2016) provide an exception, studying a small discount brokerage, but do not look at overall market effects.

investigations have concerned whether posters contribute value-relevant content.⁹ A smaller but substantial body of research has studied the use of social media as an investor relations tool that provides incremental value to news media, as in Jung et al (2015), and particularly for smaller, less visible stocks, as in Blankespoor et al (2014). A third line of study has utilized social media as a proxy for investor beliefs or expectations.¹⁰ Different from these studies, I focus on return comovement, and I make a causal statement.

The two papers closest to mine are Bailey et al. (2016) and Jiang et al. (2016). Bailey et al. (2016) exploit Facebook users' contact with far-away friends to instrument local housing demand. Jiang et al. (2016) study a message board in China and its impact on Chinese stock co-movement. My study differs heavily along methodological lines, provides significant steps toward identification, and studies a market less pre-disposed to influence by retail investors. Whereas Chinese stock market is less disciplined by institutional investors, the US stock market is the premiere equity market in the world, making it a more stringent setting to test the causal effects of social media. Also, by developing identification methods based on certain features of *modern* social media common to many platforms, I provide an example that can be adapted for future research in other contexts or for other outcomes.

This chapter also contributes to the literature on the origins of excess comovement. In a simple taxonomy of the drivers of excess comovement - institutional trading, retail investor attention and psychology, and costly information - I develop the causal link between social media and excess return correlation. While prior work such as Kumar and Lee (2006) has documented increased correlated retail trading associated with increased comovement in returns, the channel by which coordination occurs is unclear, and while consistent with causality, such prior work remains only

⁹Chen et al. (2014) find value-relevant information in *Seeking Alpha*, Jame, Johnston, Markov and Wolfe (2016) find that *Estimize* users beat the consensus I/B/E/S forecast, and Da and Huang (2015) perform a novel field experiment on *Estimize* users to show that social media posters are more informative when they prevented from herding. A large literature in computer science uses social media as a testing ground for using machine learning methods to predict stock returns. A far from exhaustive list starts with Bollen, Mao and Zheng (2011), Zhang Fuehres and Gloor (2011), Bollen and then abounds.

¹⁰Along these lines, Giannini et al. (2015) and Cookson and Niessner (2016) use StockTwits to measure investor disagreement. Further, Kwan (2015) studies StockTwits as a proxy for investor attention, finding user-reported sentiment on StockTwits is broadly uninformative and predicts future return reversal.

suggestive. This chapter argues social interaction and digital communication may be an important coordination mechanism, and moreover, that this correlated activity may in fact be *causal* on co-movement. Finally, such prior evidence uses retail brokerage account data from the early 1990s - my evidence may be somewhat surprising considering the continued growing of institutional ownership in US markets in the last two decades, and the general perception that there has been a large reduction in retail trader participation in financial markets.¹¹ Yet, while these trends might portend to the marginalization of their influence, to the extent retail traders coordinate via proliferating societal practices such as social media, it may still be possible to have a sizable gross footprint on financial markets despite thinning ranks.

1.2 Data

1.2.1 About StockTwits

Data for this study come from StockTwits, a social media website allowing users to share thoughts about stocks and market conditions in short messages, similar to its inspiration Twitter. StockTwits started in 2009 and is open to the public, inviting users to create accounts, supply a digital biography, and post content in the form of short messages containing text, videos and charts. Subject to terms of service, there are no limits to poster activity. This is in sharp contrast to *Seeking Alpha*, for example, which has an editorial board that reviews and rejects submissions and publishes on a delay.

StockTwits is a pioneer in financial social media. It was the first platform to propose social media as an investor relations tool during earnings calls. In addition, it designed its platform to enable users to quickly search for content about specific stock tickers. StockTwits invented a linguistic convention called the “cashtag”, which consists of a ticker symbol, preceded by a dollar sign (ex. \$AAPL). It is a widely used convention, which has since been replicated by competitors such as Twitter, Estimize and many others. In 2012, StockTwits extended the cashtag to include “Bullish” and “Bearish” tags, allowing users to codify their predictions or sentiments about a stock.

¹¹Please see Stambaugh (2014), AFA Presidential Address.

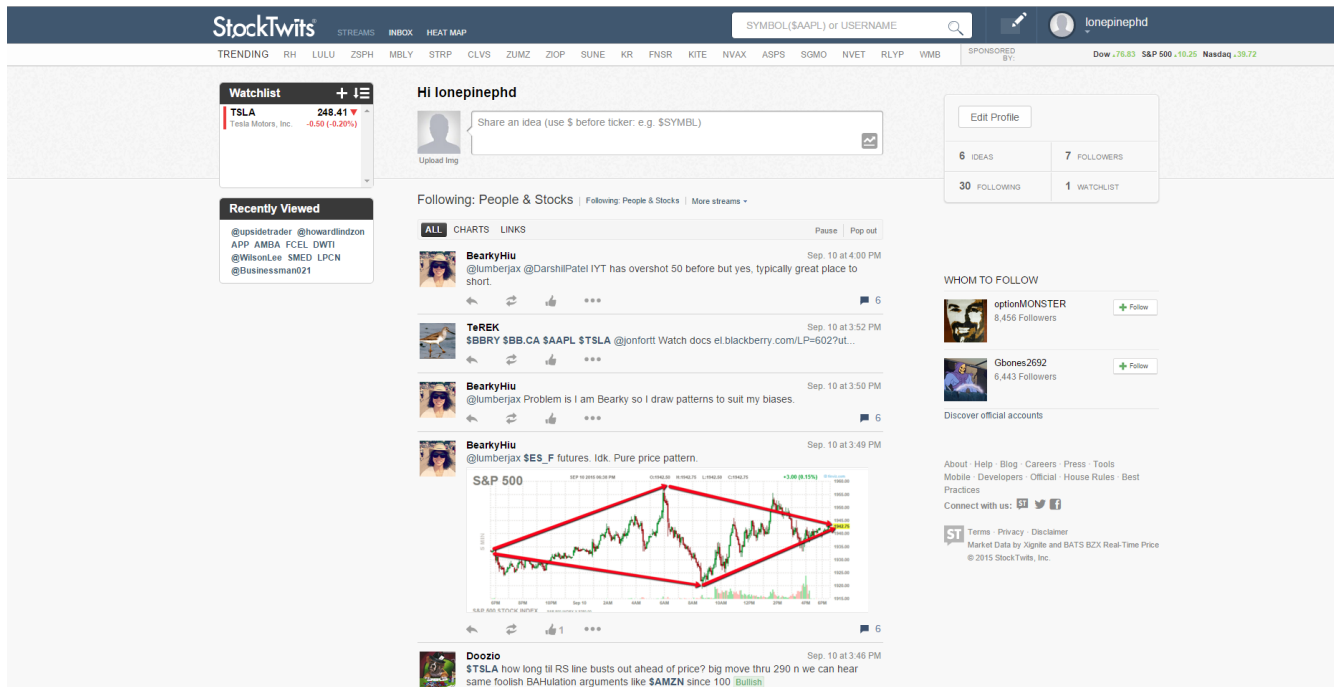


Figure 1.1: Stocktwits user interface

This is the StockTwits home page using its web interface, using the author's main account. The frontpage displays content in the form of short messages, usually coming from other users or stocks the user of StockTwits has subscribed to, or content which StockTwits' algorithms believe may be interesting to the user.

Figure 1 displays a snapshot of the StockTwits user interface through my user account. The interface of the site has changed throughout its life, but resembles the current format with two key features. Like Twitter, the center of the page is occupied by posts prioritizing stocks or posters the user “follows”, in descending order of post time. Different from Twitter, StockTwits curates specific types of content relevant to stock trading. On the top of the page, StockTwits places “Trending” tickers, used to funnel user attention into tickers of greatest interest on the site. When viewing a stock page, the sides of the screen display live market data and statistics concerning current message activity and sentiment. StockTwits does post a small number of ads. However, the site’s revenue primarily comes from licensing its content or data for use by media sites, hedge funds or companies, and fees paid by public companies who use the site for investor relations.

StockTwits today calls itself the “#1 social media site for finance”, a claim implicitly speaking to the site’s popularity. Among crowd-sourced financial media sites, its web traffic is second to the five-years-older *Seeking Alpha*. That said, its user base is sizable. StockTwits claimed a user base of 250,000 in 2012, ballooning to a million unique visitors per month in 2015 and 1.6 million in 2016. In 2015, AlexaRank ranked StockTwits in the top 20 of its category worldwide, and SimpleWeb in September 2016 ranked it #41.

These statistics do not count web traffic, for example through its mobile app, nor user impressions formed on the content redistributed through the several dozens of business partners StockTwits has. A few of these partners are boutique discount brokers such as Robinhood or Interactive Brokers or regional news media, but others are large news platforms, such as Bloomberg, Thomson Reuters, Yahoo! Finance (started in 2011, ended in 2016) and Marketwatch.com (ongoing, signed in 2014). These partnerships lead to prominent display of its content. For example, Yahoo! featured StockTwits’ content for every ticker page and Marketwatch displays the StockTwits “Trending” banner at the bottom of every page on its site. Pictures showing these integrations are displayed in Appendix B.1 and B.2. Although the site and its esteemed list of partners grows over time, unfortunately, no single partnership event stands out to represent a regime shift or interesting event study during my sample period, a conclusion echoed by StockTwits staff members I talked

to. Interestingly, in July 2016, StockTwits users can now directly trade via the discount brokerage Robinhood, which may be an interesting event study in the near future if StockTwits releases this data to academic partners.

I supplement the StockTwits dataset with data I scraped from the site directly. I first collected follower lists of every poster on StockTwits through June 2016. For each user’s webpage, StockTwits allows a social media consumer to navigate through the poster’s user list. Although StockTwits does not indicate when the poster-pair relationship was initiated, StockTwits evidently displays followers in descending order of when added. For instance, for Howard Lindzon, the founder of StockTwits, the last page of his account displays posters who whose accounts were started in 2009. The ordering of the follower list is likely a website engineering convenience as it avoids the computational expense of sorting lists and updating every server servicing StockTwits content. Given I know the order in which consumers became followers of a given poster, I am able to back out a user’s followers based on StockTwits’ reported posters as of a prior post. Additional details on how I collected the data and related details that may impact my inference will be discussed in Section 5.1, when I introduce the experiment.

I also collected high frequency snapshots of the StockTwits website, which I began collecting in late May 2015 and ended April 5, 2016. The end date was arbitrary. For the purpose of this study, I use data from complete months spanning June 2015 to March 2016. I recorded the HTML from the website every ten seconds on two different computers, giving me a complete record of the “Trending” banner. I will design an instrumental variable using this data, which I describe in 5.2.

1.2.2 Other data

For this study, I only consider US equities, defined in CRSP as SHRCD 10 or 11. A large plurality of the site’s discussion is related to ETFs, but the vast majority relate to individual stocks. Over-the-counter stocks are restricted since early on in 2010 to avoid any concerns of being used as a platform for market manipulation.

I study return correlations at an investor-relevant monthly level, as well as a short horizon (in-

trading, weekly) befitting the frenetic pace of social media. When studying returns at a monthly level, I use daily returns from CRSP. When studying returns weekly level, I use half-hour returns constructed from DTAQ, using the NBBO cleaned in a manner in keeping with the suggestions from Holden and Jacobsen (2014).¹²

I merge with StockTwits based on ticker-date, checking matches by name.¹³ Stock characteristics supplementing the main analysis are pulled from standard finance databases. I/B/E/S is used to obtain analyst coverage, Thomson-Reuters 13/F is used to retrieve institutional ownership, and firm characteristics where relevant are retrieved from Compustat. Capital IQ Key Developments are used to proxy for the arrival of corporate news events.

1.2.3 Calculating correlations

Because correlation could arise from common exposure to state variable, I purge the returns of factor exposure before computing correlations. I first calculate daily return correlations within a month, using in most months between 20-21 observations. For daily returns, this means taking residuals of 200-day rolling five-factor (Fama French 2015) time-series regressions as in Frazzini and Pedersen (2014). Some studies prefer to do beta estimations separately from the observation period of the market variable of interest, but this is not my concern as I am not predicting future returns. I require 100 daily returns to calculate a beta.

For the weekly horizon, I look at half-hour returns. This is based on 13 bars per day, 5 days a week - 65 bars. In general, when there is no price update, I carry over the mid-point price from the prior period. When the NBBO lacks a valid price to start the day, I cannot calculate a return for the first bar. If a stock is missing any two such bars throughout the week, I delete. Within each week,

¹²The focus of Holden and Jacobsen (2014) is primarily to discuss how best to clean MTAQ data. For DTAQ, which I use in this chapter, the main requirement is filter out certain items in the “quote condition” field and any internally inconsistent prices ($bid \geq ask$, for example).

¹³After removing punctuation, StockTwits’ tickers and names line up exactly with the names reported in CRSP. The management of these lists is thorough by StockTwits, as name and ticker changes match to the day either of these change in CRSP.

I winsorize across all stocks to minimize issues related to synchronicity of trading.¹⁴

My benchmark model uses SPY as a proxy. For convenience of computation, I am calculating within-week betas. That is, I calculate exposure to the market using all data within the calendar week (e.g. Monday returns are purged of betas incorporating Friday data of the same week). As a quick note on robustness, there are slight sampling differences since using 200-day rolling betas requires more data before a stock pair can enter my sample, but using the same pair-months does not alter any directional inferences.

I also compute a variety of other correlations. At the intraday level, I use the percent spread, volume, order imbalance, Amihud liquidity ratio - when no trading, I impute the value of 0 (e.g. infinite elasticity). I compute the analogous data at the daily level (except order imbalance, which lacks an analogue). This is for the final analysis which looks at comovement in market activities such as liquidity and trading volume.

1.2.4 Social media summary statistics

In this section, I describe site statistics. In Table 1.1, I present statistics about the StockTwits user base, the types of stocks that are popular on StockTwits. Panel A displays annual posting and poster counts. The site has doubled in the quantities of postings almost every year, increasing from 1 million posts in 2011 to 7 million in 2014 and 15 million in 2015. Since 2012, there are several thousand posters who post on average once daily. Note that these numbers are smaller than the relevant figures quoted in the media for “registered” users. For example, media clippings suggest 250,000 users in 2012, but my calculations suggest only 22,451 posters who ever posted in 2012. This suggests, as is common with many social media sites, many users consume but do not contribute content.

Panel B reports the number of symbols discussed at various frequencies. I refer to 2014, the middle of the sample. On any given day, the site discusses 1/4th of the equity universe (CRSP

¹⁴I have assessed the sensitivity of my inferences at the weekly horizon by simply requiring a certain number of price updates, or by winsorizing correlations (which may be excessively high when two stocks do not trade). The main results are not substantially quantitatively different.

share-codes 10,11), and 3/4ths of the equity the universe in any given month. For each column that shows the coverage of the equity universe, I also show the number that reach the threshold of 50 posts during that time interval. This suggests that while discussion is disperse, it displays some tendency to concentrate or cluster. Roughly 1 out of every 50 stocks discussed in a given week garner at least 50 posts total. In unreported tabulations, I find across the years that daily Herfindahl of posts is the range of of 1-3%, increasing to 2.8% in 2012 to 1% in 2015 – equivalent to equal share between 40-100 stocks.

Figure 2 reveals the posts are largely concentrated during the trading day. Intriguingly, the pattern of message volume follows the classic U-shape pattern documented as far back as Admati and Pfleiderer (1988), that reaches its peaks at the start and end of the trading day. It suggests that retail investor attention follows similar patterns to broader market activity.

Table 1.2 describes the distribution of posters characteristics. Panel A shows that 44% of posters who provide a self-description are “Technical” traders, on a volume-weighted basis, and 20% of posters are “Momentum” traders, categories that describe the archetypal daytrader. In terms of posters’ trading experience, over half of posters consider themselves “Intermediate” or above, suggesting that StockTwits posters are generally active in trading. Panel B describes quantiles of the StockTwits followership base. The bulk of posters have fewer than 100 followers, and so the cross-sectional distribution of followers is highly right-tailed.

In Appendix Table A.3, I describe stocks by popularity. Split by decile, stocks that garner the most message volume unsurprisingly are stocks that are “attention-grabbing” – monotonically higher beta, growth, large cap and 12-month momentum. Although not perfectly monotonic, relatively popular stocks characteristics include annual skewness, volatility and absolute return. In A.4, I report some popular names by size tercile in 2015. Among the largest stocks, consumer technology companies (Facebook, Apple, Twitter, Tesla) dominate. Among the mid-size and lowest-size tercile categories, the most popular stocks cover both consumer tech and pharmaceuticals.

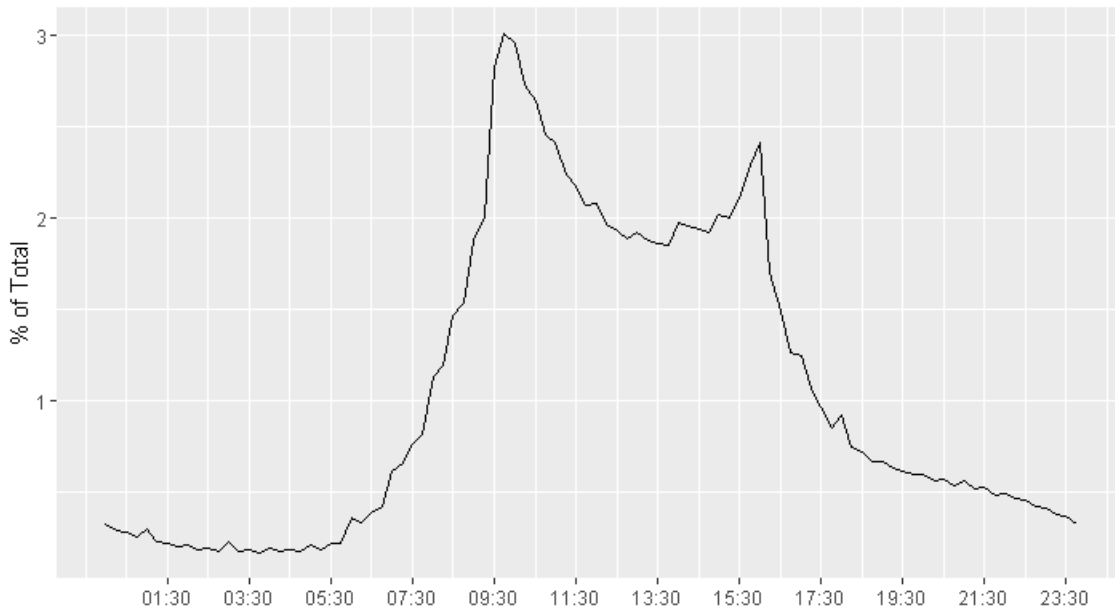


Figure 1.2: Post Distribution Throughout Trading Day

In this graph, I depict the empirical frequency distribution of StockTwits public user messages by half hour interval. On the x-axis are times of the day, on the y-axis is the percentage of all-time messages by StockTwits posts in the half-hour (e.g. 9:30-9:59:59).

1.3 Methodology

1.3.1 Empirical design

The goal of this chapter is to investigate a speculated causal relation between social media activity and comovement. To do this, I will perform panel regressions of the form:

$$\rho_{ijt}^{\epsilon} = \alpha_t + \alpha_{ij} + \beta_1 \text{SocialMediaActivity}_{ijt} + \text{controls}_{ijt} + \psi_{ijt}$$

In the above regression, ij refers to a pair of stocks i and j at time t , ρ_{ijt}^{ϵ} is the excess return correlation over the period t . α_t is the time fixed effect while α_{ij} is the pair fixed effect, ruling out the concern social media activity covers inherently correlated stocks or time periods of abnormal cross-sectional correlation. Some of the very recent studies that look at excess comovement use

Table 1.1: StockTwits and Message Volumes

Panel A: Message Volumes on StockTwits (Referring to Trading Tickers), June 2009-June2016 In this table, I present statistics about the annual number of posts on the StockTwits website, referencing a stock ticker currently trading. The variable *#Messages* describes the number of posts that occurred in that year viewable by a member of the public. *#NewPosters* describes the users that first appeared in that year. *Annualized* multiplies the number by 12*#months in my dataset for that year. *#DailyPosters* describes the number of posters who post at least 365 times.

| year | # Messages (10^6) | # New Posters | Annualized | # Users ≥ 1 post | # Daily Posters |
|------|-----------------------|---------------|------------|-----------------------|-----------------|
| 2009 | 0.400 | 3,810 | 11,430 | 3,810 | 323 |
| 2010 | 0.900 | 5,681 | 5,681 | 8,367 | 762 |
| 2011 | 1.700 | 12,095 | 12,095 | 17,189 | 1,366 |
| 2012 | 3.200 | 19,616 | 19,616 | 28,880 | 2,467 |
| 2013 | 4.200 | 21,522 | 21,522 | 35,294 | 3,047 |
| 2014 | 6.100 | 31,691 | 31,691 | 50,563 | 4,440 |
| 2015 | 8.400 | 46,891 | 46,891 | 74,798 | 5,981 |
| 2016 | 5.500 | 29,852 | 59,704 | 65,359 | 4,148 |

Panel B: Number of Stock Symbols (SHRCD 10,11) Covered Per Day, Week or Month In this table, I present statistics of the average number of US equities covered per day, week or month. I first present the number of stocks with at least 1 mention in a 1 posting. I then present average number of stocks with at least 50 postings in a day. The same statistics are calculated at weakly and monthly horizons.

| year | #Daily | $\geq 50posts$ | #Weekly | $\geq 50posts$ | #Monthly | $\geq 50posts$ |
|------|---------|----------------|---------|----------------|----------|----------------|
| 2009 | 297.7 | 2.3 | 940.5 | 7.7 | 1,551.8 | 19 |
| 2010 | 396.7 | 3.5 | 1,162.2 | 11.9 | 2,025.9 | 31.6 |
| 2011 | 614.7 | 7.4 | 1,616.3 | 24 | 2,579.8 | 63.8 |
| 2012 | 835.3 | 16.3 | 2,020.1 | 51.5 | 3,007.6 | 128.8 |
| 2013 | 1,072.1 | 22.2 | 2,515.6 | 66.1 | 3,287.7 | 162.7 |
| 2014 | 1,324.6 | 37.4 | 2,832.5 | 103.1 | 3,457.7 | 243.4 |
| 2015 | 1,657.5 | 53.1 | 3,097 | 144.9 | 3,544.5 | 336.6 |
| 2016 | 1,611.6 | 62.8 | 3,084.4 | 157.8 | 3,471.5 | 352.7 |

Panel C: Distribution of Poster "Sentiment Cashtags"

This table displays counts of posts from July 2009-June 2016 containing a "sentiment" tag, which is a user's forward looking prediction about a stock ticker. A *Bullish* tag indicates the poster is positive about the stock, *Bearish* indicates negative sentiment. The sample period begins in 2012 when StockTwits introduces the feature. The vast majority of messages contain a stock symbol, but not all contain a message concerning sentiment.

| | N | % |
|---------|------------|--------|
| Bearish | 1,666,130 | 4.170 |
| Bullish | 6,697,820 | 16.750 |
| All | 39,992,152 | 100 |

Panel D: Distribution of Posts By Trading Venue of Mentioned Symbol In this table, I list the proportion of posts that refer to tickers that trade on the listed exchange or trading venue listed below. I grouped all NYSE equity exchanges together for convenience. Foreign Exchange and INDEX refer the price of a foreign currency in the spot exchange market, or the price of an index (such as the SPX). "Miscellaneous" refers to tickers that do not relate to a traded security, but to topics generated by StockTwits.

| Exchange | %Message |
|------------------------------|----------|
| NASDAQ | 44.6 |
| NYSE | 44.4 |
| Chicago Mercantile Group | 3.8 |
| INDEX | 2.6 |
| Foreign Exchange | 2.6 |
| StockTwits "Miscellaneous" | 1.1 |
| Toronto Stock Exchange (TSX) | 0.4 |
| OTC | 0.3 |
| Private Placement | 0.1 |
| ICE | 0.1 |
| TSX Ventures | 0 |
| EUREX | 0 |

panel regressions at the pair-time level, with time fixed effects and firm characteristics used to model comovement.¹⁵ Relative to most studies on comovement, this chapter is both (a) relatively short-term, spanning four years and (b) measures correlations at a higher frequency. At these higher frequencies, firm characteristics, such as book value or institutional ownership, would be very slow-moving. But, these higher frequencies give me a great deal of within-stock variation to exploit. Thus, departing from these prior studies, I use the extra power I have due to repeated observations of stock pairs, and use within-firm, within-time approach, with standard errors double clustered by firm and time. These overloaded regressions are generally more conservative than other approaches I have surveyed from the literature - my magnitudes are larger when I use similar specifications to other papers.

How can social media generate comovement? Generally, social media may generate comovement if it generates content that induces consumers of social media to trade on some stocks, but not others. There are two ways this can occur. First, akin to analysts, posters may specialize in producing information about specific stocks. To the extent their posting influences other posters' actions, this may generate comovement if this increases the likelihood the affected subset is traded at the same time. Second, social media is unique in that it coalesces information across different contributors and directs that content to interested consumers. Social media content may percolate in a variety of ways. For example, posters may simply log-in to the StockTwits and simply observe recent posts. Another possibility would be social media posters expressing interest in a particular topic, and information about all stocks related to that topic may be gathered by social media platforms and presented in one area.

A third possibility - which I exploit in this chapter - is that social networks may be used to direct content to users that are socially connected, perhaps not just through first, but perhaps second or third order connections. Although I cannot directly measure the extent to which information is

¹⁵A non-comprehensive list of examples include Israelsen (2014), who studies correlation at the annual level, and Anton and Polk (2014), who study correlation using daily excess returns. Alternative studies, such as Green and Hwang (2009) interpret comovement as compute stock betas to the relevant index. This approach is more common when calculating the increases in comovement (β) of stocks to portfolios; this approach does not fit here. I have tried an adaptation of this approach and yielded similar inferences.

viewed and transmitted through social media, I attempt to capture this spread of information by instrumenting co-posts using posts transmitted through social networks. That is, I argue that even if the posts themselves are not exogenous, the transmission of information itself is exogenous, which may result in re-posting. If two stocks are packaged together through common followers, then this may impact *co-posting*.

To operationalize the first method, I coin the term *co-post*. Over some time horizon, I calculate the pairs of stocks a poster discusses. The horizon I choose is one hour, which is arbitrary. If social media consumers log on every few hours, this may be appropriate; if they log on once a day, and check their favorite posters' latest content, perhaps combining across the day is appropriate. Narrowing the horizon to five minutes or widening to a day does not substantially affect my results. To operationalize co-posts across a month, I aggregate all poster-hours in which a stock i and j were discussed. Specifically:

$$\sum_h^{hours} \sum_a^{posters} 1\{poster_a \text{ discussed } j \ \& \ poster_a \text{ discussed } i\}$$

At face value, this variable is not exogenous. A poster may report about two stocks that face a common event, that also drives return correlations. In 5.2, I will instrument this variable.

To operationalize the idea social media coalesces information, I coin the term “concurrent post”, to represent coalescing of information that arrives synchronously. The main idea is that two independent posters, speaking to a common audience, may act in concert about these two pieces of information. To be clear, I cannot actually observe the content viewed by social media posters, and there may be other dimensions along which a social media platform decides to send content to some audiences, but not others. However, to the extent social media posters see information at the same time, consumers of social media may see these stocks around the same time, and they may be act in concert about the two stocks. At a time interval t , I calculate the hours in which a pair of stocks was discussed in the same hour h .

$$\sum_h^{hours} \sum_a^{posters_{ih}} \sum_b^{posters_{jh}} |\#posts_i + \#posts_j|$$

Of course, the issue is that the two posters may be reacting to events occurring at the same time, causing correlations and posting to both increase. Conditional on a level of posting, I design a variable that captures the tendency for some posts to reach larger audiences due to pre-determined variation in their audience size. The idea is based on “intransitive triads”, a concept often used in the social network literature. The idea is to take every pair of *independent posters* discussing *independent stocks*. The only variation I deem exogenous is the number of followers-in-common belonging to some pairs rather than others. These followers-in-common are likelier to view these pairs of posts as opposed to other concurrent posts.

To operationalize this concept, I calculate the number of followers-in-common to all poster-pairs ab (poster a and poster b) that discuss stocks i and j where $i \neq j$ and $a \neq b$. I define the term *CommonFollowers $_{ijt}$* :

$$\sum_h^{hours} \sum_a^{posters_{ih}} \sum_b^{posters_{jh}} |followers_a \cap followers_b|$$

I also define an alternative metric, which describes the number of socially connected pairs. This provides an alternative measure which lowers the influence of extremely popular posters. I define the term *PairsCommonFollowers $_{ijt}$* :

$$\sum_h^{hours} \sum_a^{posters_{ih}} \sum_b^{posters_{jh}} 1\{|followers_a \cap followers_b| > 0\}$$

Because common followership is pre-determined, this variable allows me to make a causal statement. I further describe the experimental design and particulars of the common-follower calculations using this measure in 5.1.

1.3.2 Sample selection

One limiting factor in this study is the inability to study the entire panel of stock pairs over a long time period.¹⁶ However, given that my my setting, StockTwits, and social media more broadly, grows in size dramatically over my sample, I should study social media over a relatively long time period. Thus, to strike a balance, I select a tractable 360,000 stock pair subset and track from a time period spanning 2012 to March 2016.

I narrow my sample to a subset of pairs where *some* threshold level of social media activity occurred over the sample. The threshold I chose is arbitrary, but narrows the sample substantially. After doing so, I pick a subset of stock pairs that looks most like the broader stock universe. In this way, in other words, I choose among stock pairs with social media activity in a way that looks as representative of the broader stock universe as possible. The steps are as follows. First, I filter all pairs that receive a certain threshold level of monthly co-postings (5) in a given month at least once in my sample. Specifically, I choose any stock pair that has ever received a co-posting, e.g. a poster in the same hour talks about both stock i and stock j .¹⁷ This yields approximately 1 million stock pairs. The number still is still intractably high. For each i and j , I compute its median stock size bin over the sample period. Within each size quartile pair (small-small, small-medium, small high, etc.) I randomly sample 22,500 pairs, yielding 360,000 pairs over the whole sample. Appendix A.1 displays the sample selection process steps and the observations at each step.

How does my sampling procedure affect my inferences? First, by looking at only stock pairs with any social media activity, I limit my inferences to intensive margin estimates. To some extent, extensive margin estimates are difficult as my site has at best a large plurality of social media

¹⁶Consider a 4200 stock universe, roughly the number of stocks in CRSP at the end of 2014. The lower diagonal of the correlation matrix is $4200 \times 4199 / 2 = 8817900$ cells. Without any changes to the equity universe, a monthly panel over a year would be 10.5 million observations; a week panel over a year would be 45.85 million observations; over a year a daily study would be 2.222 billion observations. Each regression would contain at least 8.81 million fixed effects, with around 4200 more for every additional stock added to the sample. These numbers are dramatic underestimates as there are in fact 13 million possible pairs since 2011 due to stock listings and delistings, business combinations, etc. This is without considering the computational problems posed by processing the correlation matrices itself, as well as DTAQ, copost measures and common follower measures.

¹⁷Many alternative selection methods are possible. Although perhaps one alternative would be to use the concurrent-posts as a filter, the original draft of this chapter used this co-posting filter.

Table 1.2: Cross-sectional Distribution of Posters' Characteristics
 Panel A: By Most Recently Self-Reported Classifications

In these tables, I report the number of total posts from June 2009-June 2016 by StockTwits users falling the reported category. The categories are self-defined, and designations are taken as of a posters' most recent post by June 2016. The first panel presents self-categorized "Experience". A blank indicates no response. The second panel presents self-designated trading classification. A "Day Trader" refers to positions that last intraday. A "Swing Trader" refers to positions lasting between 2 days to 2 weeks. A "Position Trader" holds stocks at horizons longer than that. The third table shows trading approaches. Momentum refers to strategies exploiting recent return movements. Technical strategies include Momentum strategies, but refer to rule-based strategies that incorporate return movements and security fundamentals. Global Macro refers to asset-class level strategies taking into account macroeconomic conditions.

| Experience | # Users | # Message | % |
|--------------|---------|------------|--------|
| | 102,417 | 14,195,842 | 35.498 |
| Novice | 22,271 | 2,932,289 | 7.332 |
| Intermediate | 32,462 | 10,489,743 | 26.231 |
| Professional | 13,998 | 12,372,674 | 30.939 |

| Holding Period | # Users | # Message | % |
|-----------------|---------|------------|--------|
| | 101,991 | 14,532,962 | 40.529 |
| Day Trader | 14,647 | 5,536,592 | 15.440 |
| Swing Trader | 27,246 | 10,815,786 | 30.163 |
| Position Trader | 14,114 | 4,972,867 | 13.868 |

| Trading Approach | # Users | # Message | % |
|------------------|---------|------------|--------|
| | 104,079 | 14,803,240 | 37.017 |
| Technical | 26,507 | 12,253,475 | 30.641 |
| Momentum | 12,062 | 4,325,606 | 10.817 |
| Global Macro | 2,286 | 923,253 | 2.309 |
| Growth | 11,163 | 2,387,820 | 5.971 |
| Fundamental | 8,601 | 3,094,580 | 7.738 |
| Value | 6,450 | 2,202,574 | 5.508 |

Panel B: By Followership (as of Latest Post)

In these tables, I report the number of total posts from June 2009-June 2016 by StockTwits users falling the reported category. The categories are mutually exclusive categories of the size of the poster’s following. A following is the number of users that have decided to subscribe to a poster’s messages, and the site will prioritize delivery of content from these posters to the social media user above others. Posters’ follower counts are taken as posters’ most recent post by June 2016.

| Follower Range | % | # Posts | # Users |
|--------------------------------------|-------|------------|---------|
| [0-9) | 24.69 | 9,873,849 | 3,007 |
| [10-100) | 14.02 | 5,605,323 | 144,579 |
| [100-1000) | 31.57 | 12,626,133 | 22,488 |
| [1000-10 ⁴) | 16.12 | 6,445,851 | 517 |
| [10 ⁵ , 10 ⁶) | 11.91 | 4,764,070 | 553 |
| [10 ⁶ , ∞) | 1.69 | 676,926 | 14 |

activity. Second, one might wonder if characteristic balancing introduces any bias. I argue the purpose of doing so is to remove this bias, but in my original draft, I did not balance by size and the inferences were identical. However, in the cross-section, it was difficult to draw inferences because characteristics were imbalanced.

1.3.3 Sample summary statistics

In this section, I define variables and describe summary statistics of the sample. Appendix A.1 defines all variables again for convenience. Table 1.3 reports the summary statistics.

Panel A of Table 1.3 reports statistics for the monthly sample. The mean excess correlation is 0.93%, with a large standard deviation of 24.09%. The average raw return correlation has a similar standard deviation, but is much higher on average: 24.5% for this sample. I will discuss the measures of social media activity in the relevant sections.

Next, I discuss the measures of social media activity. $\log(1+\#CommonFollowers_{ijt})$ approximates the number of “common impressions” to the stock pair. For this sample of stocks, the standard deviation is 3.49, or about 32 common followers, and the 25th to 75th percentile range is about

Table 1.3: Sample summary statistics

This table reports summary statistics at the pair-time level (pair-month, pair week). N_{ijt} is the number of institutional owners owning an above median share of stocks and as of the prior quarter. ρ_{ijt}^{ret} and ρ_{ijt}^{ff5} are the correlations of residual returns against the relevant benchmark. ρ_{ijt}^{ff5} refers to residuals of CRSP daily returns purged of 200-day-beta exposure to the Fama-French Five Factors; ρ_{ijt}^{ret} refers to residuals of returns purged of intra-week-beta to the SPY ETF. σ_{ijt} refers to the correlation of absolute returns. ρ_{ijt}^{raw} refers to correlation of raw returns; CRSP daily in the monthly panel. $\log(1+\#Coposts_{ijt})$ is the log count of poster-hours in which a poster discussed both i and j at some point in the hour, aggregated across the relevant horizon. The hours are aligned so that when not during trading hours, the statistic contributes to the next trading day. $\log(1+\#PairsCommonFollowers_{ijt})$ is defined as the log count of hour-poster pairs where a poster is discussing stock i and another poster is discussing stock j in the same hour (or time unit otherwise specified), where the two posters have a common follower and the poster has at least 50 followers. $\log(1+\#AboveMedianCoowners_{ijt})$ is the same, except each pair is scaled by the number of followers-in-common. The common followers are calculated from at least six months prior to the week or month of posting. τ_{it}^{month} is the number of posts about stock i during period t . Returns and turnover are from CRSP. Extended descriptions are provided in Appendix A.3 and Section 3.4.

Panel A: Monthly

| | N | μ | σ | 1st | 5th | 25th | 50th | 60th | 75th | 80th | 95th | 99th |
|--|-----------|-------|----------|--------|--------|--------|-------|-------|-------|-------|-------|-------|
| $\log(1+\#AboveMedianCoowners_{ijt})$ | 7,064,231 | 2.10 | 1.89 | 0.00 | 0.00 | 0.00 | 2.64 | 3.22 | 3.76 | 3.93 | 4.86 | 5.34 |
| ρ_{ijt}^{ff5} | 7,062,343 | 0.93 | 24.09 | -53.96 | -38.21 | -15.39 | 0.74 | 6.84 | 17.03 | 21.09 | 40.71 | 57.94 |
| ρ_{ijt}^{ret} | 7,062,297 | 24.53 | 26.46 | -39.38 | -20.39 | 6.59 | 25.48 | 32.44 | 43.67 | 47.93 | 66.23 | 78.53 |
| $\log(1+\#CommonFollowers_{ijt})$ | 7,064,231 | 3.66 | 3.49 | 0.00 | 0.00 | 0.00 | 3.40 | 4.64 | 6.31 | 6.93 | 9.98 | 11.72 |
| $\log(1+\#PairsCommonFollowers_{ijt})$ | 7,064,231 | 1.59 | 1.69 | 0.00 | 0.00 | 0.00 | 1.10 | 1.61 | 2.64 | 3.00 | 4.90 | 6.50 |
| $\log(1+\#Coposts_{ijt})$ | 7,064,231 | 0.36 | 0.63 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 0.69 | 1.79 | 2.56 |
| τ_{it}^{month} | 7,064,048 | 1.17 | 15.71 | -34.97 | -18.13 | -4.88 | 1.03 | 2.98 | 6.47 | 8.05 | 19.49 | 41.77 |
| Turnover _{it} | 7,064,231 | 2.67 | 11.07 | 0.21 | 0.56 | 1.13 | 1.76 | 2.12 | 2.93 | 3.37 | 6.98 | 13.87 |

Panel B: Weekly

| | N | μ | σ | 1st | 5th | 25th | 50th | 60th | 75th | 80th | 95th | 99th |
|---|------------|-------|----------|--------|--------|-------|-------|-------|-------|-------|-------|-------|
| $\log(1+\#AboveMedianCoowners_{ijt})$ | 42,935,297 | 0.12 | 0.69 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.11 |
| $\rho_{ijt}^{ret,SPY}$ | 42,935,104 | 3.06 | 20.12 | -46.76 | -28.83 | -9.08 | 2.61 | 7.04 | 14.92 | 18.24 | 36.20 | 56.05 |
| ρ_{ijt}^{ret} | 42,935,104 | 16.12 | 21.79 | -39.28 | -17.86 | 2.06 | 10.03 | 15.32 | 20.84 | 30.36 | 34.20 | 67.05 |
| $\log(1+\#CommonFollowers_{ijt})$ | 42,935,297 | 2.00 | 2.90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 3.87 | 4.71 | 8.13 | 10.57 |
| $\log(1+\#SociallyConnectedPoster_{ijt})$ | 42,935,297 | 0.81 | 1.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 1.39 | 1.61 | 3.50 | 5.14 |
| $\log(1+\#Coposts_{ijt})$ | 42,935,297 | 0.13 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 1.79 |
| τ_{it}^{week} | 42,933,993 | 0.23 | 6.71 | -17.07 | -8.30 | -2.32 | 0.21 | 1.09 | 2.65 | 3.36 | 8.51 | 18.22 |
| Turnover _{it} | 42,933,993 | 13.47 | 30.31 | 1.97 | 3.09 | 5.56 | 8.74 | 10.53 | 14.61 | 16.77 | 36.59 | 76.75 |

630 common followers. I next display the summary statistics for $\log(1 + \#PairsCommonFollowers_{ijt})$, which counts not the number of common followers, but posters with common followers. Some posters with large followings may dominate the $\log(1 + \#CommonFollowers_{ijt})$ statistic, so this provides an alternative measure. Naturally, its standard deviation is smaller as the connection is not multiplied by the number of followers. The standard deviation is about 5 socially connected posters, while the interquartile range going from 0 to 20.

To the extent the plausibility of this hypothesis depends on a large audience size, it is important to note that these numbers underestimate the economic magnitude of these audiences to the extent the content is redistributed through sub-followers, or organically through “click-throughs”, the act of a website user following a link to another page. For example, if consumer Andrew follows poster A and B, and consumer Scott follows Andrew, then Scott may see Andrew’s feed by clicking on Andrew’s profile link. In that way, Scott’s followers may see Andrew and his feed, and all such posters may see A and B. Click-throughs can happen in an almost endless number of ways, such as through sites well-integrated with StockTwits such as Yahoo! or Marketwatch. A final caveat is that since I am counting posts in a narrow interval, the actual number of “concurrent impressions” may be much higher over the course of a long day. For example, over the course of a day, poster A may post at 9AM, but poster B may post at 10AM. A consumer making a choice at 12:00PM may see both of these posts then.

I next discuss my measure co-posts. This variable has a standard deviation of 0.63, or about 2 coposts. The same caveats apply: these magnitudes generally underestimate the number of actual impressions, to the extent the content is circulated to a large audience, and the co-posts are counted in a narrow interval.

Panel B reports the weekly sample. All summary statistics are sensible. The magnitudes of social media activity are smaller (smaller time interval over which to aggregate), and the average excess correlation is higher (fewer factors in the market model). The only sharply different statistic is the number of common owners; 200 day rolling betas requires 200 days since an IPO. I check this difference is not inference-changing.

1.4 Baseline correlations

1.4.1 Baseline panel regression

In this section, I discuss a baseline result, that co-postings positively relate to excess return correlation. These coefficients are meant to assess economic magnitudes, leaving identification to Section 5.

In Columns 1-4, I establish the relation between co-postings and correlation. I will use the baseline specification as follows:

$$\rho_{ijt}^{\epsilon} = \alpha_t + \alpha_{ij} + \beta_1 \log(1 + \text{Coposts}_{ijt}) + \beta_2 \log(1 + \#\text{Posts}_{it}) + \beta_3 \log(1 + \#\text{Posts}_{jt}) + \text{controls}_{ijt} + \psi_{ijt}$$

The logic of the specification is that the number of co-posts is correlated with the number of posts idiosyncratic to i and j , and thus capture all stock-specific posting. Column 1 shows that at the monthly level, direct co-postings about pairs of stocks are highly significantly related to contemporaneous factor-adjusted return correlation. The effect is within-pair, within-day. Column 2 repeats the exercise at the weekly level. In Column 1, the very precisely estimated coefficient of 0.7545 ($t=13.35$) on $\log(1 + \#\text{co-postings})$ suggests that otherwise equal, co-postings are associated with increased factor-adjusted correlation. It means that going from zero to 3 posts is associated with 75.45 basis points of excess return correlation which is economically large in general. Coefficients on $\log(1 + \#\text{Posts}_{it}), \log(1 + \#\text{Posts}_{jt})$ are negative, suggesting that stock-specific posting, proxying perhaps for stock-specific events on either i or j , attenuates the pairs' comovement, which is sensible.

In Column 2, the coefficient for intraday half-hour returns at a weekly horizon is smaller 0.345, suggesting the magnitudes at a weekly level are smaller but still economically large. The same pattern on $\log(1 + \#\text{Posts}_{it}), \log(1 + \#\text{Posts}_{jt})$ appears, which is that otherwise equal, idiosyncratic movements in either stock attenuates comovement. That the coefficient on weekly returns is smaller is not surprising for a number of reasons. It is possible, for example, that microstructure

Table 1.4: Baseline correlations

The variable of interest is $\log(1 + Coposts_{ijt})$, which is log count of poster-hours over the course of the time period that posted about stock pair ij . Additional detail about these variables and additional variable definitions are provided in Appendix A.2. All regressions, as indicated, contain pair-time (week or month) fixed effects. The sample period for the monthly sample is 2012-March 2016; for the weekly sample, it is October 2012 to March 2016. Robust standard errors double-clustered by pair, time period are reported in parentheses.

| | ρ_{ijt}^ϵ | | | |
|-------------------------------------|-----------------------|-----------------------|----------------------|-----------------------|
| | Monthly | Weekly | Monthly | Weekly |
| $\log(1+Coposts_{ijt})$ | 0.755*** (0.057) | 0.349*** (0.036) | 0.751*** (0.057) | 0.352*** (0.035) |
| $\log(1+Posts_{it})$ | -0.120*** (0.022) | -0.373*** (0.023) | -0.118*** (0.023) | -0.353*** (0.025) |
| $\log(1+Posts_{jt})$ | -0.101*** (0.021) | -0.340*** (0.0218) | -0.097*** (0.023) | -0.309*** (0.023) |
| ret_{it} | | | -0.002 (0.002) | -0.015*** (0.005) |
| ret_{jt} | | | -0.003 (0.002) | -0.013*** (0.005) |
| $ ret_{it} $ | | | 0.0017 (0.002) | 0.0040 (0.008) |
| $ ret_{jt} $ | | | 0.000 (0.002) | -0.001 (0.009) |
| $turnover_{it}$ | | | -0.001 (0.001) | -0.000 (0.001) |
| $turnover_{jt}$ | | | 0.001 (0.002) | -0.001 (0.001) |
| $\log(1+AboveMedianCoowners_{ijt})$ | | | 0.032 (0.029) | 0.0886*** (0.034) |
| $ret_{it} * ret_{jt}$ | | | 0.001*** (0.0002) | 0.009*** (0.0012) |
| $ ret_{it} * ret_{jt} $ | | | -0.0002* (0.000) | -0.0030*** (0.001) |
| $turnover_{it} * turnover_{jt}$ | | | -0.000 (0.000) | 0.000 (0.000) |
| Pair FE? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| Time FE? | <i>Month</i> | <i>Week</i> | <i>Month</i> | <i>Week</i> |
| Num. obs. | 7,062,343 | 42,935,104 | 7,061,890 | 42,933,800 |
| Adj. R ² | 0.0505 | 0.0795 | 0.0506 | 0.0802 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

noise contaminates the estimate. Conversely, it is possible that social media posters may influence stock prices over the course of a day, and relatively less adept to handling information that arrives at a faster pace such as a half-hour interval.

Turning to Columns 3 and 4, I present specifications that include controls for returns, absolute returns, turnover and above median owners. In both monthly and weekly estimates, the main effect is resilient, weakening by half a percent at the monthly level, and strengthening by 2% at the weekly level. The first three controls simply look for concurrently-timed market events, which might mechanically generate correlation. Above-median owners accounts for the an albeit relatively low-frequency change in the number of institutional co-owners, which can exacerbate correlation. The estimate is resilient to these controls, with minor loss of observations stemming from missing data in CRSP (usually for missing volume). The main takeaway is that the correlation is not due to simply return events at the same time. I look for other types of events in the next section. The coefficient on above median owners is positive, but insignificant at the monthly level, but interestingly significant at the weekly level.

1.4.2 Robustness - is social media activity a substitute for news?

I perform a variety of robustness checks, some of which I report on the Online Appendix. In this section, I address the concern that social media is just a substitute for news, such as timed news releases, or industry events. Such events may gather attention, yielding trading and excess return correlation, with social media activity reporting on these events as a byproduct. In this analysis, I cast doubt on that concern by evaluating the relation between co-posting and excess return correlation during periods of news event arrivals.

To define the news, I rely on event taxonomy of Capital IQ Key Developments. Capital IQ Key Developments is a database that is used by investment bankers to track events, and provides a list of announcements, categorized by Capital IQ into categories and filtered for duplicates, and the announcement date-time. As discussed in Edmans et al. (2014), CapitalIQ discovers 234 different events, and the database's event coverage predates my sample period. From this taxonomy of

Table 1.5: Robustness - Co-postings during News Events

In this table, I report panel regression estimates of the form:

$$\rho_{ijt}^{\epsilon} = \alpha_t + \alpha_{ij} + \beta_1 \log(1 + Coposts_{ijt}) + \beta_2 \log(1 + \#Posts_{it}) + \beta_3 \log(1 + \#Posts_{jt}) + controls_{ijt} + \psi_{ijt}$$

The variable of interest is $\log(1 + Coposts_{ijt})$, which is log count of poster-hours over the course of the time period that posted about stock pair ij . Earnings, Bankruptcy and Business Expansion (and Product Announcement) events are defined as those from Capital IQ's Key Developments database. The $industry_i - industry_j - time$ fixed effects are based on the Fama French 49 affiliation of stock i and stock j coalesced from {Compustat where available, CRSP where available, 49 otherwise}. The controls are $turnover$, $\log(1 + \#AboveMedianCoOwners)$, ret , and $|ret|$. Additional detail about these variables and additional variable definitions are provided in Appenndix A.2. All regressions have the indicated fixed effects. Robust standard errors double-clustered by pair, time period are reported in parentheses.

| | ρ_{ijt}^{ϵ} | | | | | | |
|---|-------------------------|----------------------|---------------------|--------------------------|----------------------|---------------------|---------------------|
| Horizon: | Earnings Month | Earnings Week | Bankruptcy Month | Business Expand Month | AnyKeyDev Month | Industry Month | Industry Week |
| $\log(1+Coposts_{ijt})$ | 1.072*** (0.102) | 0.607*** (0.061) | 1.247*** (0.335) | 0.739*** (0.058) | 0.908*** (0.080) | 0.730*** (0.053) | 0.290*** (0.028) |
| D_i^{event} | -0.211*** (0.061) | -0.110*** (0.038) | -0.253* (0.130) | 0.068 (0.072) | 0.001 (0.0421) | | |
| D_j^{event} | -0.216*** (0.052) | -0.116*** (0.038) | -0.197 (0.135) | 0.010 (0.085) | 0.020 (0.040) | | |
| $\log(1+Coposts_{ijt}) * D_i^{event}$ | 0.001 (0.113) | -0.217*** (0.063) | -0.325 (0.338) | -0.093 (0.091) | -0.237*** (0.061) | | |
| $\log(1+Coposts_{ijt}) * D_j^{event}$ | 0.122 (0.111) | -0.227*** (0.064) | -0.212 (0.315) | -0.010 (0.076) | -0.220*** (0.065) | | |
| $D_i^{event} * D_j^{event}$ | 0.306*** (0.079) | 0.014 (0.041) | 0.143 (0.127) | 0.177 (0.320) | 0.067 (0.052) | | |
| $\log(1+Coposts_{ijt}) * D_i^{event} * D_j^{event}$ | -0.212 (0.146) | -0.061 (0.075) | -0.008 (0.338) | -0.485** (0.227) | 0.085 (0.080) | | |
| Time FE? | Yes | Yes | Yes | Yes | Yes | $Ind_i * Ind_j * t$ | $Ind_i * Ind_j * t$ |
| Pair FE? | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Num. obs. | 7,064,231 | 42,935,297 | 7,064,231 | 7,064,231 | 7,064,231 | 7,064,231 | 42,933,800 |
| Adj. R ² | 0.0446 | 0.0802 | 0.0507 | 0.0506 | 0.0507 | 0.0705 | 0.1060 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

events, I extract several types of events: earnings events, industry events, corporate expansions, bankruptcies.¹⁸ I create dummy indicators that ask whether news events occur in a given month or in the two weeks straddling a week. The specification I run captures the arrival of an event on stock i , stock j , and both stock i and stock j . Specifically:

$$\rho_{ijt}^{\epsilon} = \alpha_t + \alpha_{ij} + \beta_1 \log(1 + Coposts_{ijt}) + \beta_2 \log(1 + \#Coposts_{ijt}) * 1\{Event_{it}\} + \beta_3 \log(1 + \#Coposts_{ijt}) * 1\{Event_{jt}\} + \beta_4 \log(1 + \#Coposts_{ijt}) * 1\{Event_{it}\} * 1\{Event_{jt}\} + controls_{ijt} + \psi_{ijt}$$

Table 1.5 reports the results. Models 1 and 2 explore the behavior of the point estimates during earnings months. I perform a triple interaction term that captures the relation between posting and correlation when either or both of the stocks involved is during an earnings month/week. Column 1 and 2 show the behavior of correlations during monthly and weekly time periods. Two patterns emerge from the data. First, otherwise equal, an earnings event for a company will tend to decrease the correlation between two stocks; when they coincide, correlation tends to increase, although the latter effect is weak for the weekly horizon. Second, the relation between concurrent earnings announcements and co-postings is insignificantly negative, suggesting that co-postings are less likely to be associated with correlation during earnings events. Meanwhile, the unconditional coefficient on co-postings, β_1 , increases dramatically, from 0.75 to 1.072 at the weekly horizon and from .35 to .61 at the monthly horizon, increases of 40% and 80%, respectively. The simplest interpretation is that earnings events are periods of highly idiosyncratic news and return movements. Outside of these exogenous events, all else equal, the unconditional effect of co-postings and correlation becomes much stronger. Similarly, Columns 3, 4 and 5 report business expansions, distress, or any type of Cap IQ Key Development. The conclusion is similar. If anything, controlling for exogenous news events, the unconditional relation between social media activity is stronger.

The other obvious explanation to any time-variation in correlation is an industry-level fundamental or an industry-level increase in trading activity. To rule this out, in columns 6 and 7, I add industry-pair-time fixed effects. Industry is based on the Fama-French 49 industries, where indus-

¹⁸The events that Capital IQ tracks are often times granular events within a broader class of events I consider. For example, Capital IQ tracks both a “Dividend Initiation” and a “Preferred Dividend”, an “Earnings Announcement” as well as the announcement of a future earnings announcement. I categorize the list of events by type and aggregate them together.

try affiliation is determined by the SIC code reported in Compustat, then in CRSP, if not available in Compustat (49 default). These capture all week/month events common to industry pairs, defined by Fama French industries. The number of fixed effects increases the R-squared by 50% in the weekly regression and 80% in the monthly regression, suggesting this is a large increase in model terms (2401 per period for 49*49 industry pairs). Yet the main point estimate is largely resilient, dropping about 2 basis points at the monthly level and 6 basis points at the weekly level. Of course, it remains possible that my industry classification is imperfect.

Taken together, this suggests social media activity is on average weaker during news events, not stronger. A news substitution story would yield the opposite. I have tried many additional robustness checks. My analysis is robust to re-defining co-posting at different time subsamples, and different baseline control specification.

1.4.3 Robustness - timing of posts

One concern is the nature of relating contemporaneous posts to contemporaneous returns. Although posting could cause same-day comovements, that then manifest in monthly returns, there may be reaction with a delay. Therefore, we should expect that posts from the first half of a month affect posts from the second half, and so forth. Appendix A.7 explores this possibility. It decomposes co-posts into halves or quarters, and then measures the intensity of co-posting in relation to posting. The first two columns show that posting from the first half of the month have larger coefficients than the second half. Similarly, columns 3 and 4 show posts from the first week matter more than posts from the last week. However, the effects are not strongly well-ordered, with posts from the second week driving comovement more so than others. However, the last week is notably the least important economically.

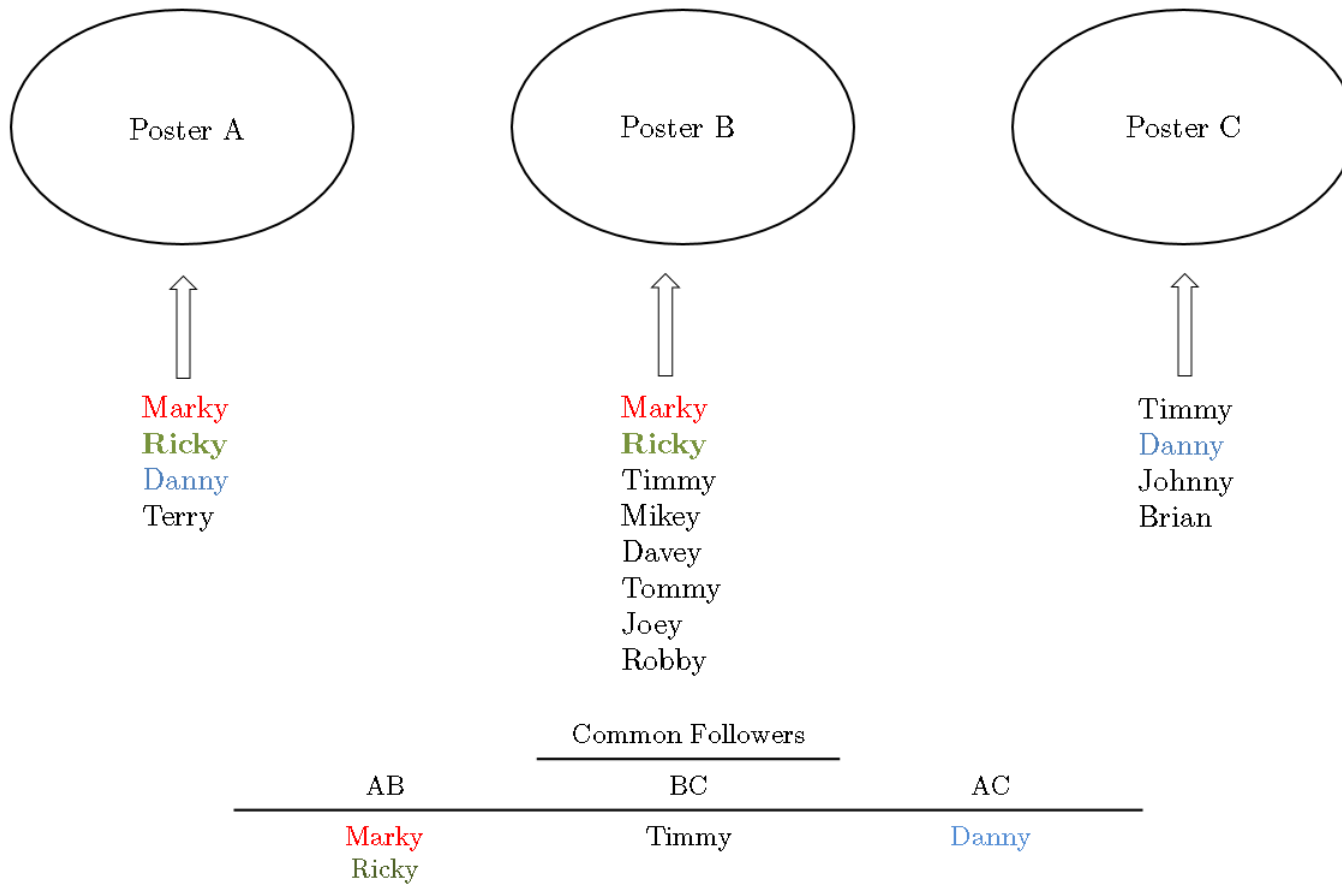


Figure 1.3: Sample Common Follower Calculation For Identification Strategy

1.5 Common follower networks as an instrumental variable

1.5.1 Design

In my first experiment, I argue causality by exploiting how information is transmitted through social networks. Social media platforms provide a platform to distribute content and communications and have algorithms for determining how to distribute that information. One principal way these decisions are made is through social networks. Forming social networks on a social media platform is implicitly subscribing to a social media user's content. Take three posters discussing three different stocks at roughly the same time. Denote them poster A, B, C, discussing stocks X, Y, Z. A viewer of social media that follows A and B will be more likely to observe the postings on X and Y than she would Z. Based on this, said viewer would in turn trade and/or post on X and Y as well, in turn generating more trading.

I operationalize this idea in an instrumental variable design. The IV is some variant of the count of independent poster pairs that share a common follower. Because the count of common followers requires posting to occur, I control for the number of concurrent posts, regardless of common followership, because otherwise my IV is endogenous to the omitted variable also driving posting. First, I create the term $\log(1 + \text{WeightedConcurrentPosts}_{ijt})$ which simply counts the number of posters about stock i , multiplied by her followers, plus the number of stock j , multiplied by her followers, occurring in the same hour, summed across the month. I perform a large variety of specification checks and present a view in foregoing analysis. Next, after counting the total level of posting, I count the common followers to the posts for all such posts that occur in the same hour. For each stock pair, I aggregate the number of common followers, or simply count the number of poster pairs that are socially connected. Again, I only count the followers if the followership decision was made at least six months prior to the event. The first stage is therefore:

$$\log(1 + \text{Coposts}_{ijt}) = \kappa_1 \log(1 + \text{CommonFollowers}_{ijt}) + \kappa_2 \log(1 + \text{WeightedConcurrentPosts}_{ijt}) + \kappa_3 \log(1 + \#\text{Posts}_{it}) + \kappa_4 \log(1 + \#\text{Posts}_{jt}) + \text{controls}_{ijt} + \alpha_t + \alpha_{ij} + \phi_{ijt}$$

The second stage is therefore:

$$\rho_{ijt} = \beta_1 \log(1 + \widehat{Coposts}_{ijt}) + \beta_2 \log(1 + \text{WeightedConcurrentPosts}_{ijt}) + \beta_3 \log(1 + \#Posts_{it}) + \beta_4 \log(1 + \#Posts_{jt}) + \text{controls}_{ijt} + \alpha_t + \alpha_{ij} + \psi_{ijt}$$

The quantity of interest in this example is β_1 , which is an estimate of the effect of common mentioning that is due to the extent of how often two stocks are presented to the same social media user. For two stocks with the same level of posting, the pre-determined common followers can be interpreted as the effect of social media in transmitting news. Because the common followers were based on a social network determined before the event, it rules out reverse causality concerns. The other concern is omitted variables. As I control for the level of posting, and the level of posting can be interpreted as a proxy for poster reaction to outside news. Thus, I interpret β_1 as a measure of the effect of social media's transmission of content, controlling for a level of news.

1.5.2 Social graph data

To run the exercise above, I need to the social “graph” of every user. StockTwits does not provide this data, despite academic inquiry by a number of researchers. However, as explained in 3.2, I infer the historical follower lists as of time t from the data StockTwits provides me, which lists the number of followers as of the time of a post (or when StockTwits recorded a post). I then scrape the StockTwits website, which lists followers in the order they were added. The reason they list followers in the order they were added is probably that this method is simple and requires no additional effort or computation time to do otherwise. Using these two numbers, I can back out the likely follower lists, assuming that survivorship bias is not severe. To the extent that I am lagging posts by six months, the bias is likely very low.¹⁹ Also, the six month lag mitigates any issue in the early part of the sample with respect to how followers were recorded.²⁰

I calculate the following quantities. First, as of every month, for computational reasons, I filter

¹⁹To the extent survivorship bias would exist, it would be that a great number of followers deleted their accounts or removed chose not to follow a poster. As I am using a lagged value of followers, I am unlikely to be counting anyone who wasn't a follower at time t , but I may be missing some posters who were followers but later chose to remove themselves. The other type of survivorship bias would be that posters no longer exist. Of over 81,000 posters by June 2016, fewer than 100 no longer had accounts - most users that ceased to post rarely deleted their accounts. For either type of survivorship bias, I would take the stance this would simply be measurement noise.

²⁰The lag of post recording is not usually longer than 1 week, or in extreme cases 1 month, but I check the main inference is robustly significant in the final year, in which no asynchronous recording of followers occurs.

out posters who six months ago did not have at least 50 followers. Then, I calculate for every such poster their followers-in-common with every other poster in the network and then convert it to a count. This list is extremely large, with the symmetric list reaching 134 million pairs of posters with followers in common by June 2016, from 81,234 posters who have 50 followers or more. This means that 1.5% of all possible pairs of posters with 50 followers or more have a common follower. Appendix C plots time series of the follower network, listing the number of total followers, as well as the edge counts, over time.

From these edge counts, I then compute the common followers as described in Section 3.4. To recap, for a given month and its six-month-lagged common follower network, I loop across all hours. Across all hours, I consider all poster-stock/poster-stock pairs, where the posters and stocks are not the same (e.g. pairs of different posters discussing different stocks). For each poster, I record the totally hourly counts of the number of socially connected pairs of posters $\log(1+\#PairsCommonFollowers)$, and the total of their common followers $\log(1+CommonFollowers)$.²¹

1.5.3 Results

Table 1.6 reports my results. Panel A presents results using posting and correlations at a monthly resolution, while Panel B presents results at the weekly resolution. The main point of emphasis will be the monthly results, with the weekly results to match the pace of social media.

Turning to Panel A, I find that the relation between instrumented co-posting and return correlation to be positive. Interpreting column 1, I find that sensibly, the greater the number of firm-specific posts, $Posts_{it}$ and $Posts_{jt}$, the greater the number of coposts. This must be at least somewhat true by construction, as co-posts increase both of these quantities. Interestingly, $ConcurrentPosts_{ijt}$ is negatively related to co-posts. An albeit ad hoc interpretation may be that during times of large

²¹This is simply the sum of the followers shared by posters A and B, and posters B and C, etc. If a follower common to A and B is also common to B and C, I double count. To the extent this consumer receives two impressions of a pair of stocks, this is an adequate measure. However, one might conjecture that this follower should only be counted once. However I cannot, for computational reasons, store and load the lists of common followers and take the union of all followers/content consumers in common for a stock given pair. The computational complexity of such a problem (not considering the space required to store this data), is $O(n^2 * t * p^2 * s^2)$, where p is the number of posters, t is the number of time bars, and n is the number of followers per poster p , for all stocks s . Keeping only the number of common connections, it is $O(t * p^2 * s^2)$.

Table 1.6: Instrumental variable based on on common follower networks

This table presents instrumental variable regressions. The endogenous regressor is $\log(1 + \widehat{Coposts}_{ijt})$. The instrument is the number of common followers, $\log(1 + CommonFollowers)_{ijt}$, or the number of $\log(1 + \#PairsCommonFollowers)_{ijt}$. Panel A presents monthly results, Panel B weekly results. $\log(1 + Posts_{it})$ is the number of posts on stock i throughout the month t . $\log(1 + CoincidentTotalPosts)_{ijt}$ is the number of posts in the same hour (or 15 minute window, in the case of weekly intraday results) regardless of whether the posts have common connections. $\log(1 + CoincidentTotalPosts * Followers)_{ijt}$ is the same, except each post is weighted by the number of followers of the posters. Where polynomial order controls are used, the terms $\log(1 + CoincidentTotalPosts * Followers)_{ijt}$ and $\log(1 + CoincidentTotalPosts)_{ijt}$ are added to the specification up to the specified order. All regressions are estimated with pair and time fixed effects, with cluster robust standard errors at the pair-time level.
 Panel A: Monthly

| | | ρ_{ijt}^{FF5} | | | | |
|--|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| $\log(1 + \widehat{Coposts}_{ijt})$ | | 1.580*** (0.214) | | 1.476*** (0.261) | 2.178*** (0.301) | 2.626*** (0.304) |
| $\log(1 + CoincidentTotalPosts * Followers)_{ijt}$ | -0.009*** (0.001) | 0.090*** (0.010) | -0.020*** (0.002) | 0.088*** (0.011) | .061 (0.027) | -0.241*** (0.054) |
| $\log(1 + CoincidentTotalPosts)_{ijt}$ | | | | | 0.811*** (0.085) | -0.311** (0.157) |
| $\log(1 + Posts_{it})$ | 0.034*** (0.005) | -0.342*** (0.035) | 0.093*** (0.008) | -0.325*** (0.041) | -0.467*** (0.039) | -0.557*** (0.041) |
| $\log(1 + Posts_{jt})$ | 0.030*** (0.005) | -0.322*** (0.037) | 0.094*** (0.008) | -0.305*** (0.044) | -0.443*** (0.043) | -0.525*** (0.046) |
| $\log(1 + PairsCommonFollowers)_{ijt}$ | 0.204*** (0.009) | | | | | |
| $\log(1 + CommonFollowers)_{ijt}$ | | | 0.062*** (0.003) | | | |
| Num. obs. | 7,060,175 | 7,060,175 | 7,060,175 | 7,060,175 | 7,060,175 | 7,060,175 |
| F-stat | | 34.34 | | 32.11 | 18.66 | 13.055 |
| FE? | Yes | Yes | Yes | Yes | Yes | Yes |
| IV? | | Pairs | | CommonFollowers | Pairs | Pairs |
| Control polynomial order? | 1 | 1 | 1 | 1 | 2 | 3 |
| Controls? | No | Yes | Yes | Yes | Yes | Yes |
| Adj. R ² | 0.476 | 0.050 | 0.453 | 0.050 | 0.050 | 0.050 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Panel B: Weekly

| | ρ_{ijt}^{SPY} | | | | | |
|--|--------------------|------------|------------|-----------------|------------|------------|
| $\log(1 + \widehat{Coposts}_{ijt})$ | | 0.875*** | | 0.320 | 0.673*** | 1.421*** |
| | | (0.192) | | (0.200) | (0.174) | (0.229) |
| $\log(1 + \text{CoincidentTotalPosts} * \text{Followers})_{ijt}$ | -0.004*** | 0.019*** | -0.006*** | 0.015*** | .1309 | -0.079** |
| | (0.001) | (0.005) | (0.000) | (0.005) | (.020) | (0.040) |
| $\log(1 + \text{CoincidentTotalPosts})_{ijt}$ | | | | | 0.134*** | -0.455*** |
| | | | | | (0.058) | (0.099) |
| $\log(1 + \text{Posts}_{it})$ | 0.047*** | -0.432*** | 0.064*** | -0.357*** | -0.458*** | -0.485*** |
| | (0.004) | (0.028) | (0.003) | (0.030) | (0.028) | (0.031) |
| $\log(1 + \text{Posts}_{jt})$ | 0.047*** | -0.390*** | 0.064*** | -0.314*** | -0.417*** | -0.444*** |
| | (0.004) | (0.027) | (0.003) | (0.030) | (0.028) | (0.032) |
| $\log(1 + \text{PairsCommonFollowers}_{ijt})$ | 0.133*** | | | | | |
| | (0.005) | | | | | |
| $\log(1 + \text{CommonFollowers}_{ijt})$ | | | 0.032*** | | | |
| | | | (0.001) | | | |
| FE? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | Yes | Yes | Yes | Yes | Yes | Yes |
| F-stat | | 54.22 | | 72.29 | 35.82 | 22.49 |
| IV? | | Pairs | | CommonFollowers | Pairs | Pairs |
| Control polynomial order? | 1 | 1 | 1 | 1 | 2 | 3 |
| Num. obs. | 42,933,800 | 29,004,813 | 42,933,800 | 42,933,800 | 42,933,800 | 42,933,800 |
| Adj. R ² | 0.340 | 0.080 | 0.320 | 0.080 | 0.080 | 0.080 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

firm-specific news, posters are less likely to discuss pairs of stocks, but more likely to discuss firm-specific news. However, the coefficient on the variable of interest, $PairsCommonFollowers_{ijt}$, referring to the number of posters posting in the same hour connected by at least one common follower, is positive. This suggests that when stocks are juxtaposed together, there is an increased tendency to co-mention. That is, it is not that simultaneous events matter, but simultaneous *viewership* and the posting it generates, that impacts return correlation. The F-statistic of 34.34 is well above the necessary condition recommended by Stock and Yogo (2002). In the second stage, I find that indeed the instrumented co-posting is strongly related to comovement. A standard deviation increase (0.69) increases comovement by about 1%, double the unconditional mean and is 5% of the standard deviation. This magnitude is large and sensible. $ConcurrentPosts_{ijt}$ is also positively related to comovement. Simultaneous posting may represent simultaneous events, which may be endogenous to an omitted variable which drives coincident movements in prices.

Columns 3 and 4 present alternative instruments based on counting the common followers, not the intransitively connected pairs. The estimates are quantitatively very similar. Columns 5 and 6 present specification checks. One potential concern with this strategy is that even if I control for the concurrent posting, β_1 is significant because of the correlation between followership and concurrent posting. I mitigate this concern in two ways. First, in the forthcoming analysis, I present the results of raising concurrent posts to the 2nd and third order. Doing so strengthens the economic magnitude of my results. Second, re-specifying the analysis as a ratio between followers and the level of concurrent posting yields similar results, although the first-stage F-statistic is not reliable. $ConcurrentPosts_{ijt}$ and $ConcurrentPosts*Followers_{ijt}$ turn positive in Column 5 (polynomial order 2) but negative in Column 6 (polynomial order 3), meanwhile unreported coefficients on the polynomial terms are positive.

Panel B reports weekly results. Note that the clock for counting coincident posts and common followers is now 15 minutes, given the return bars are now based on a window of 30 minutes. The coefficient estimates are less reliable and smaller in magnitude, but are in most specifications significant. My preferred estimate from Column 2 suggests that a standard deviation movement in

co-posting (.37) is related to 87.5 basis points of comovement, or .375 for a one standard deviation movement. Column 4 presents results based on common followers. The estimate is positive and economically large, but insignificant. The instrumental variable is very reliable, but heteroskedasticity in the count of common followers might attenuate the coefficient of β_1 . However, counting the number of socially connected posters appears to work well, and as columns 5 and 6 indicate, appears robust to polynomial specification checks.

1.5.4 Endogeneity concerns regarding social connections

Social connections may be endogenous. I assume that many choices of followership are idiosyncratic in nature, at least with respect to whether or not a pair of stocks is discussed by two independent posters at a future point in time. However, it is possible that social connections embed information about an event unrelated to social media, which may drive both posting and comovement. This would suggest followership decisions are made because followers of posters decide to follow posters who track events exposed to a common future event. That is, it is possible followership decisions are made because they are correlated with events that cause high *conditional* correlation. For example, if social media users follow gold stock posters, I may simply be capturing a persistent shock to gold stocks.

I argue this is unlikely to be driving my results. First, specifications that take into account recent changes to the pair's correlation mitigate the possibility that the followership networks are correlated with an event that also causes high conditional correlation. Specifically, controlling for the lagged level in correlation does little to affect my main result. In addition, pair-year fixed effects reduce the possibility any local change in pair relationship drives both the relation between followership and co-posting and the return correlation.

Third, and most convincingly, Appendix A.6 presents statistics on what percentage of the stock universe a poster tends to cover, and a tendency for that poster to cover the same stock or industry ever again. A couple of interesting facts are revealed. First, the average poster covers about 41 symbols, and weighted by the number of intransitive connections they have in the social network

(number of other posters with which they share a common follower), the average number is 301, covering 9 sectors and 123 of roughly 200 StockTwits industries. This suggests that no industry event could explain my findings. As we saw also in Table 5, my results are robust to industry-by-industry-by-time fixed effects and robust when not considering stocks in the same Fama-French 49 industry. It also suggests that given the breadth of coverage, any followership decision is unlikely due to be systematically tracking a particular type of stock, because the range of stocks covered by any poster is so large. The final panel of Appendix A.6 suggests that the conditional probability a poster will ever mention the same stock, ever again, is less than a quarter. For two posters covering two different stocks independently, the probability is 1%. Thus, any recurring, common event that explains followership and posting must reconcile with the fact that the that recurring, common event very rarely results in a poster-pair referring to the same stock pair, at the same time, in the future.

1.6 Instrumental variable based on StockTwits banner

1.6.1 Exclusion restriction

In this section, I propose an instrumental variable approach. My work in Section 5.1 obtained identification using the idea that quasi-random, predetermined variation in social media common followership created an audience common two stocks. These common audiences process this information more heavily than other information presented, and the covered subset of the stock universe comoves more as investors trade. In this section, instead of social media's role in passing through information, I argue that social media can also guide or shape content production.

To operationalize this concept, I use the StockTwits "Trending" banner. The idea is that posters are likelier to discuss stocks on the banner, all else equal, because the banner directs their attention to a particular set of stocks. I am going to use both the presence of a stock on the banner as an instrument, as well as *where* it is on the banner. The "Trending" banner is every page of StockTwits. In addition, variants are displayed on the sites of many partners that integrate its

content. Appendix B.1 shows that Yahoo! Finance displayed the top 10 “Trending” stocks in many places of its site. Appendix B.2 shows that Marketwatch integrates StockTwits directly into its site, displaying the top 5 “Trending” stocks on every page.

To explain the exclusion restriction, it is worthwhile to describe the mechanics of this ticker banner. Every 5 minutes, StockTwits sorts stocks by message volume, transformed by a function that takes into account the 24 hour message volume and stock-specific parameters. The stock-specific parameter is not known to the public or even partners such as Marketwatch, but I was told in 2014 that StockTwits scores the stock’s message volume relative to some stock-specific measure of posting intensity in the past 9 months. After measuring the transformed message volume of all stocks on the site, StockTwits elects the top 30 and displays a certain fraction on the banner, depending on the user interface used by social media consumer.

The consequence of this stock-specific parameterization is that the stocks vary widely in their requirements to enter the StockTwits banner. For example, some stocks can enter the StockTwits banner with as few as 1 or 2 messages, while others may take dozens or even hundreds on a given day. The most popular stock, for example, is Apple Inc. with 5% of message volume, \$AAPL is on the banner during my sample only 0.5% of the time.

I present two designs. First, co-display of stocks is random, conditional on the same stock-level information. The exclusion restriction is plausible. Given a large news event, such as a sharp absolute or directional return movement, high turnover, or high message volume, posters may react in such manner so as to generate addition message volume that increases the probability the stock will enter the “Trending” banner. However, some stocks have high predetermined requirements to enter the StockTwits banner, and some have relatively low requirements. Further, for any two pairs of stocks, their propensity to be on the StockTwits banner and their visual placement relative to one another is affected by the behavior of other stocks whose message volume scores are also being transformed by an arbitrary statistical function, introducing an additional element of randomness to the exercise. Thus, subject to controls, I argue entering the StockTwits banner is effectively quasi-random. The second design is to argue that even of co-display is an endogenous event,

the conditional visual arrangement of the stocks is quasi-random and there is enough information on how stocks on the banner are visually displayed to influence the co-posting intensity. To the extent the exclusion restriction is not completely exhaustive of all alternatives, at the very least, this instrumental variable exercise limits the possible set of alternatives.

1.6.2 Data, sample period, and design

For this experiment, I have 10 months of data. I started collecting the data in mid-May 2015, and stopped in April 2016. The end date was arbitrary. As a result, I have 10 full months beginning June 2015 and ending March 2016. The data consists of snapshots of the StockTwits frontpage occurring every 10 seconds, on two different computers. This is a much higher frequency than the frequency at which the banner updates, so synchronicity issues in recording the banner are relatively minimal.

From this data, I can extract how long a stock pair was on the “Trending” banner at the same time, and *where* the stock pair is on the banner. From this information, I create three metrics: first, I calculate the fraction of the day a given stock pair is on the StockTwits banner. The expectation is that when stocks are on the board longer, there is a greater increase in the number of co-posts about them. Second, because it may be relevant *where* stocks are, both in relation to one another or whether they are in a visually salient place on the board, I calculate two measures. First, I calculate the average rank of the pair on the board. A rank of 1 means one of the stocks is in the first slot on the board, a rank of two means they are in slot 2. The stocks are ordered in terms of their message volume score from left to right. Second, I also compute the average visual distance of the two stocks when on the board. When a stock is, on average, in the #2 slot, and another stock is, on average, in the #15 slot, the distance is 13, while its rank is 8.5. The idea is that stocks that are spaced relatively close together may be relatively salient compared to other stocks, and so posters are more likely to be stoked into discussing stocks that are visually near one another, and less so when they are not. The prediction is that when stocks are further away, they are less likely to be discussed together. These three variables (time on board, conditional distance, conditional rank)

are the basis for my instrumental variables.

Finally, to the extent social media posters only respond to the social media site itself during certain hours of the day, I limit the calculation period of when stocks are on the banner from 4am to 4pm. It is relatively less likely that posters are responding to stocks on the banner during times when view posters are contributing at all.

1.6.3 Results

The results presented here are from four parsimonious designs of my instrumental variable exercise. I only report certain coefficients of interest, but report the return, turnover, volume terms to the Online Appendix for the interested reader. In the Online Appendix, I also report issues related to variable construction, summary statistics, results from other instrumental variable designs, which take into account additional IVs in the first stage, as well as some diagnostics I performed on the StockTwits banner. For example, the within-pair, within-time tendency to co-post about two stocks does depend on the relative visual distance between the two stocks when on the banner. Finally, as a minor detail, in this analysis I change the definition of a co-post to deal with subtleties of this exercise.²² The key difference is now a co-post is a mention of stocks i and j in the *same message* as opposed to the *same hour*. I do this because I want to only look at the postings that are relatively likely to be attributed specifically to presence on the StockTwits banner.

Moving to Table 1.7, I report four sets of columns corresponding to four empirical designs. The first empirical design is a fraction, which describes the percentage of the day a stock is on the board. This variable takes into account no information about the placement of the stocks relative to one another, yet it is highly relevant in the first stage (17.64), and marginal at the 10% level in the second stage. The coefficient of .4091 is 5 basis points larger than the OLS estimate, even in spite of more controls. Next, I add another instrument; and the average pairwise, visual distance

²²In prior analysis, I would define a co-post as an hour interval in which a user would describe a stock i and j , whether in the same or different posts. Now, I define a co-post as a co-post in the same message. The main reason is that as stocks may flash on or off the banner every 5 minutes, it would be less ideal to consider as a co-post any stocks that maybe were not synchronously on the board at the same time. Under the hourly co-post definition, two stocks need not be on the board at the same time to be influenced by the banner. The results are generally similar under either definition, and this is actually a somewhat stingier choice.

Table 1.7: Instrumental variable regressions at the weekly horizon using the StockTwits banner

In this table, I report results of an instrumental variable regression over the time period June 2015-March 2016, for which I have data. The instrumental variable is based on the StockTwits "Trending" banner, which places 30 stocks at the top of the StockTwits website. The first stage dependent variable is $\log(1 + \widehat{Coposts}_{ijt}^{SameMessage})$, which in this table is the number of messages about i and j across the week. The instrument is designed four different ways. The variable $FractDayOnBoard$ is the fraction of the day the stock pair is on the board at the same time. The variable $Distance|Board$ is the number of visual slots in between the stock pair when on the board. $PairRank$ is the average position of the stock pair on the board when on the board at the same time; a lower value indicates further to the left of the screen. When a stock pair is not on the board, its distance or rank evaluate to the worst possible value (30). All regressions include pair-week fixed effects. Robust standard errors double clustered by pair and week are reported in parentheses.

| Instrument design is.. | Fraction of Day Trending | | Fraction of Day Trending, Conditional Visual Distance | | Fract. Day Trending times Visual Distance | | PairRank Board FractDay in 2nd Stage | |
|-------------------------------------|--------------------------|----------------------|---|----------------------|---|----------------------|--|----------------------|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| $\log(1 + \widehat{Coposts}_{ijt})$ | | 0.409* (0.227) | | 0.436** (0.205) | | 0.468** (0.220) | | 0.884** (0.418) |
| $FractDayOnBoard_{ijt}$ | 7.199*** (0.365) | | 5.319*** (0.312) | | | | 4.465*** (0.312) | -3.418 (3.385) |
| $Distance * FractdayOnBoard_{ijt}$ | | | | | 0.617*** (0.0370) | | | |
| $Distance Board$ | | | -0.008*** (0.001) | | | | | |
| $PairRank_{ijt} Board$ | | | | | | | -0.0123*** (0.0007) | |
| $\log(1 + post_{it})$ | 0.128*** (0.005) | -0.490*** (0.080) | 0.1240*** (0.005) | -0.494*** (0.083) | 0.129*** (0.005) | -0.499*** (0.080) | 0.123*** (0.005) | -0.551*** (0.106) |
| $\log(1 + post_{jt})$ | 0.124*** (0.006) | -0.427*** (0.079) | 0.120*** (0.005) | -0.430*** (0.082) | 0.125*** (0.005) | -0.435*** (0.079) | 0.118*** (0.005) | -0.485*** (0.104) |
| FE? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| F? | | 17.64 | | 22.83 | | 12.83 | | 12.08 |
| Num. obs. | 10,009,957 | 10,009,957 | 10,009,957 | 10,009,957 | 10,009,957 | 10,009,957 | 10,009,957 | 10,009,957 |
| Adj. R2 | 0.3705 | 0.1364 | 0.3727 | 0.1364 | 0.3691 | 0.1364 | 0.3731 | 0.1363 |

***p < 0.01, **p < 0.05, *p < 0.1

of the stocks when on the board. The F-Stat improves significantly, and the second stage estimate is more precise: .5113, which is now significant at the 5% level and almost at the 1% level. The negative coefficient on distance is exactly as expected: when stock pairs are off the board, or further apart, they are not co-posted about as often. The overidentification J-Statistic is 2.14, which does not trigger any concerns for overidentification. For alternative specifications, I arbitrarily create a product of the two, which I report in Columns 5 and 6. The interpretation of .6173 is plausible as a larger distance would reduce the intensity of co-posting, while a larger exposure-time would increase it. Thus the coefficient is positive, but smaller than in Column 1, which is sensible. This works reliably well.

In the last two columns, I revise the assumption behind the exclusion restriction. The prior three regressions assumed that if subject to a variety of controls, being on the board is *exogenous*. In the last two specifications, I present an alternative argument. Under the assumption being on the board is an endogenous event, *where* the pair is on the board is plausibly exogenous. Thus, in the last instrumental variable design, I promote being “on the board” to the second stage. Column 7 shows that even after doing so, it appears *where* the stocks are on the board is relevant for co-posting (F-stat of 12.08), within-pair, within-time, after controls. The coefficient of the time the pair is on the board interacts with the coefficient on co-posting to enlarge it to 0.884. This large jump is interesting because it may suggest my prior estimates about the role of social media are underestimated. The rationale is simple. Social media posters are likely to discuss stock pairs when their components are newsworthy. News events, all else equal, attenuate correlations, as evidenced in Table 5. Commonalities in coverage on social media may serve to reverse this effect, but without correcting for endogeneity, the actual point estimate may be attenuated significantly. Accounting for such an idiosyncratic event pushes the estimated magnitude upward. This argument is compatible with the findings of Table 5, when I find that the release of corporate news events actually diminishes the social media-comovement relation as opposed to capturing it.

In summary, I provide a diverse set of evidence that suggests social media mediates the discussions of its users. To the extent it does so, it may create commonalities in discussion across stocks,

and this commonality attributable to social media may impact return comovement. While the evidence is only moderately reliable, these reflect middle-of-the-road estimates, with some specifications capable of bringing the result down to insignificant (always positive, but only reliable at the 15 or 20% level), or some specifications using additional instruments bringing the power up but increasing the risk of over-fitting (even though the overidentification J-statistic thresholds are not triggered) or an inexplicable interpretation for certain first-stage variables.

1.7 Other tests

In this section, I perform additional tests that supplement the identification strategies presented above. If social media indeed does cause asset price movement, in the cross-section, I should observe larger effects based on the characteristics of posters, stocks and time-variation in investor-reported sentiment. Finally, I explore the relation between social media activity and commonalities in trading activity and liquidity.

1.7.1 Followership

I first decompose posters by their followership. Consistent with causality, we would expect influential posters to have a larger coefficient on their co-postings. To perform this experiment, every month, I split the universe into their observed followership as of the prior month. I then perform three slices: above 500 followers, above 1000, and above the 75th percentile. Of the followers in these two groups, I select an equal amount of posters (1/4th of the smaller group) ranked by their posting activity.²³ That is, I compare the most active posters in either group to each other. I then aggregate their co-posts for that time period (weekly or monthly) and then horse race these two posting counts.

Table 1.8 reports the results. Both at the weekly and monthly level, the results consistently show that posters with greater ex-ante followings are associated with greater increases in correlation. A

²³One alternative method would be to take all posters in the lower group and all posters in the higher group, but one potential criticism might be that I am comparing the sums of different amounts of random variables, which may have undesirable statistical properties. In particular, based on the Central Limit Theorem, this may attenuate the estimate of the smaller group relative to the larger group.

Table 1.8: Do posters with more followers have greater influence?

In this table, I report panel regressions of the relation between co-posting activity and excess return correlation. In this analysis, I segment posters by their ex-ante audience size, proxied for by followership. Every month, I sort the poster universe of active posters by their end-of-month followership into two groups - above the cross-sectional 75th percentile, above 500 followers, or above 1000 followers. Of the two groups, I take the N most active (where N is 25% of the smaller sub-group), such that both groups are equal in the number of social media contributors. I then calculate aggregate co-posts across these groups.

The variables of interest are the relative sizes of the two coefficients between low and high. Specifications include firm pair - time fixed effects, and controls for returns, absolute returns, turnover and institutional co-owners. Robust standard errors double-clustered by pair and time are reported.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| $\log(1+\text{Coposts}_{\geq 75\text{thPercentile}}^{\text{low}})$ | 0.341*** (0.067) | | | 0.780*** (0.119) | | |
| $\log(1+\text{Coposts}_{\geq 75\text{thPercentile}}^{\text{high}})$ | 0.281*** (0.035) | | | 0.589*** (0.066) | | |
| $\log(1+\text{Coposts}_{\geq 500\text{Followers}}^{\text{low}})$ | | 0.130*** (0.043) | | | 0.190*** (0.072) | |
| $\log(1+\text{Coposts}_{\geq 500\text{Followers}}^{\text{high}})$ | | 0.217*** (0.048) | | | 0.671*** (0.072) | |
| $\log(1+\text{Coposts}_{\geq 1000\text{Followers}}^{\text{low}})$ | | | 0.082** (0.042) | | | 0.227*** (0.071) |
| $\log(1+\text{Coposts}_{\geq 1000\text{Followers}}^{\text{high}})$ | | | 0.224*** (0.056) | | | 0.606*** (0.082) |
| FE? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| Controls? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| Sample? | <i>Weekly</i> | <i>Weekly</i> | <i>Weekly</i> | <i>Monthly</i> | <i>Monthly</i> | <i>Monthly</i> |
| Num. obs. | 43,002,446 | 43,002,446 | 43,002,446 | 7,064,231 | 7,064,231 | 7,064,231 |
| Adj. R ² | 0.0802 | 0.0802 | 0.0802 | 0.0501 | 0.0500 | 0.0500 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

single post by a poster with more followers, otherwise equal, corresponds to a greater increase in correlation. The raw coefficient for posters above the 75th percentile is somewhat smaller than those below the 75th percentile, but much larger when standardized. Specifically, the above 75th percentile group has twice the standard deviation at the monthly level, and 1.9 times the standard deviation at the weekly level. For the 500/1000 follower splits, the standard deviations of co-postings are very close and the coefficient differences can be interpreted as report; the effect of high follower posters is 2-3 times larger than otherwise active, low followership posters. Taken

together, my evidence supports the hypothesis that co-postings bear stronger relation when posters have larger audiences.

1.7.2 Stock characteristics

I next take into consideration stock characteristics. Consistent with the notion that social media users consume social media, and trade in a manner that impacts stock returns, I expect that to observe a greater intensive margin among stocks that are more likely to be impacted by retail or noise traders in the sense of Llorente et al. (2002) and Baker and Wurgler (2006). In addition, I expected a greater intensive margin among stocks that might fit into “style classes” in the sense of Barberis and Shleifer (2003). To the extent investors search for information, and tend to acquire information about and trade stocks within these categories, we expect social media to be particularly potent *within* these style classes, as opposed to outside of them. I find evidence supporting both of these hypotheses.

Table 1.9 reports the results of splitting the sample along the lines of Llorente et al. (2002). The stocks i and j are each sorted into low, medium and high bins, creating six categories {low-low, low-medium, low-high, medium-medium, medium-high, high-high}. The breakpoints are chosen to minimize the concentration of the sample in any one of the six bins, although similar results are obtained using the unconditional sample’s 33rd and 66th percentile. However, due to power issues, certain estimates from categories, for some measures, yield unreliable point estimates. Table 1.9 suggests that the relation between co-postings and correlation is higher down the diagonal (moving from low-low to medium-medium), or going from left to right (moving from low-low to low-medium). Seven of these eight cases are monotonically ordered. The coefficients are such that the lowest group by analyst coverage, (1-%Spread), ownership or market cap are two to three times as large as for the highest group by the same measure. The general conclusion that stock pairs with low coverage, high spreads, low market cap, and low ownership obtains. Seven of eight cases are monotonic, and all corners (ll vs lh or hh) are well-ordered.

In robustness checks, I have also tried weekly sorts. In contrast to the monthly results, the ev-

Table 1.9: Is there a larger effect on high arbitrage cost stocks?

In this table, I explore the hypothesis that stock pairs with low analyst coverage, institutional ownership and market cap, or high percentage spreads, exhibit a higher relation between social media activity and stock return correlation. I present two panels, defining social media activity differently. Panel A reports results of regressions where social media activity is defined as co-posting. Panel B defines social media activity as the number of socially connected pairs between two stocks posting in the same hour, as in Table 6. For each sub-panel, I sort stock legs by one of these four characteristics, defined as of either the prior month (cap and spread), the prior quarter (ownership) or the prior year (coverage). Decile breakpoints are chosen to minimize the Herfindahl Index between the six cells. Each regression is re-run separately with a set of controls from the main specification in Table 4, with pair-month fixed effects. Significance is based on robust standard errors double clustered by pair and time.

| Analyst Coverage | | | | % Spread | | | |
|-------------------------|-------------|-------------|-------------|-----------------------|-------------|-------------|-------------|
| | l | m | h | | l | m | h |
| l | 1.848 (***) | 1.260 (***) | 0.643 (***) | l | 0.715 (***) | 0.555 (***) | 0.524 (***) |
| m | | 0.967 (***) | 0.636 (***) | m | | 1.52 (***) | 2.397 (***) |
| h | | | 0.686 (***) | h | | | 2.783 (***) |
| Institutional Ownership | | | | Market capitalization | | | |
| | l | m | h | | l | m | h |
| l | 1.675 (***) | 1.115 (***) | 0.985 (***) | l | 2.473 (***) | 1.567 (***) | 0.386 (***) |
| m | | 0.629 (***) | 0.708 (***) | m | | 0.906 (***) | 0.545 (***) |
| h | | | 0.848 (***) | h | | | 0.801 (***) |

Table 1.10: Investor style classes

In this table, I evaluate how the posting-correlation relationship when stocks are in the same "style" class. The dependent variable is FF5-residual return correlation. This is a month-stock panel. The first column tests whether the relation between co-posting and correlation is higher within an industry, with industry-time fixed effects. The remainder columns test whether the difference in characteristic decile abates the co-posting-return relation. The variable of interest is the interaction term $|bin_i - bin_j| * \log(1 + \#Coposts)$, which takes a larger value (extreme values are 0 and 9) the greater the difference in decile. Controls on returns, absolute returns, turnovers and above-median institutional co-owners are suppressed. Except for column 1, pair and time fixed effects are used. Robust standard errors double-clustered by pair and date are reported.

| | Industry | PRC | MOM | Value | HML | Invest/Asset | CMA | MKT BETA | DIV YIELD |
|---|---------------------|----------------------|----------------------|---------------------|---------------------|----------------------|---------------------|----------------------|----------------------|
| $\log(1 + Coposts_{ijt})$ | 0.697*** (0.055) | 0.960*** (0.078) | 1.082*** (0.092) | 0.759*** (0.064) | 0.877*** (0.082) | 1.083*** (0.078) | 0.873*** (0.084) | 0.910*** (0.076) | 0.908*** (0.074) |
| $\log(1 + Coposts_{ijt}) * D_{SameFF49}$ | 0.304*** (0.112) | | | | | | | | |
| $ bin_i - bin_j $ | | -0.171*** (0.031) | -0.236*** (0.019) | -0.023 (0.016) | -0.038** (0.018) | 0.035** (0.015) | -0.037** (0.017) | -0.042* (0.025) | 0.010 (0.015) |
| $\frac{bin_i + bin_j}{2}$ | | -0.369*** (0.079) | -0.118*** (0.019) | 0.166*** (0.055) | 0.008 (0.024) | -0.015 (0.025) | -0.026* (0.016) | -0.006 (0.019) | 0.024 (0.029) |
| $ bin_i - bin_j \log(1 + Coposts_{ijt})$ | | -0.064*** (0.011) | -0.098*** (0.018) | 0.001 (0.011) | -0.035** (0.015) | -0.114*** (0.020) | -0.033** (0.015) | -0.060*** (0.021) | -0.051*** (0.011) |
| Pair FE | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Time FE | Ind-Ind-Month | Month | Month | Month | Month | Month | Month | Month | Month |
| Controls? | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Num. obs. | 6,984,741 | 7,059,506 | 6,880,220 | 5,061,314 | 6,644,445 | 6,858,428 | 6,644,445 | 6,984,562 | 7,059,513 |
| Adj. R ² | 0.071 | 0.051 | 0.052 | 0.051 | 0.052 | 0.051 | 0.052 | 0.051 | 0.051 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

idence is only weakly supportive. Two of the four tabulations behave consistently: the relation between social media activity and a stocks' institutional ownership and market capitalization behave similarly to the monthly results, but sorts on percentage spread and analyst coverage do not. What explains the difference? While my monthly results are consistent with prior literature that measures marginal effects on daily returns, it may be that weekly results contain a certain amount of noise, or that there is an additional confounding variable: factors we expect to impact arbitrage costs may impact intraday price formation in a way daily returns are not affected.²⁴

Next, I consider style-classes. The style classes that investors use are inherently unobservable and may potentially change over time. I conjecture a few based on prior literature. Following Barberis and Shleifer (2003), I explore whether value, momentum and industry are relevant investor classes. Interestingly, StockTwits' "Trading Approach" field for user profiles also lists "Value" and "Momentum" as trading styles, lending credence to this choice. I also consider stocks sorted by other Fama-French Five Factor characteristics, and stocks by their price decile, following Green and Hwang (2009).²⁵ Finally, I design a dividend style class. This is based on prior work suggesting that dividends are preferred by retail investors, for instance, as in Graham and Kumar (2006).

Table 1.10 reports the results for the relation between co-posting and whether stocks are in the same style class. The analysis is at the monthly level, but all weekly results obtain. Column 1 considers the style class to be the Fama French 49 industry. The baseline coefficient of 0.69689 is about 9% smaller than the baseline coefficient. When two firms are in the same industry, the coefficient increases by 30 basis points. This coefficient is obtained controlling for industry*industry*date fixed effects, suggesting that comovement is stronger within an industry conditional on social media posting, rather than merely coinciding with an event causing it.

For the other columns in this table, I specify characteristic differences as the difference in the numbered decile. For example, if a stock is in the 10th decile of momentum and the other member

²⁴A potential explanation is that the sorting characteristics, such as market cap, spreads, institutional ownership, etc. impact the stock's microstructure, and therefore intraday price formation, in a way daily returns are not affected.

²⁵Green and Hwang (2009) find that low price stocks and high price stocks tend to comove more with stocks in their respective price deciles, suggesting that investors approximate stocks' market capitalization using their nominal share price.

of the pair is in the 8th decile, the difference is 2.²⁶ The first of these style classes I consider is nominal share price. The remainder columns discuss the five Fama-French factors as well as momentum. For the difference-in-decile regressions, I adopt the following specification:

$$\rho_{ijt}^{\epsilon} = \beta_1 \log(1 + Coposts_{ijt}) + \beta_2 |bin_i - bin_j| * \log(1 + \#Coposts) + \beta_3 |bin_i - bin_j| + \beta_4 \frac{|bin_i + bin_j|}{2} + controls_{ijt} + \alpha_t + \alpha_{ij} + \psi_{ijt}$$

The estimate of β_2 and β_4 controls for the unconditional difference and level of stock pair style. Column 2 shows that being in a different price decile reduces the effect of co-posting. The coefficient suggests that a difference of 10 deciles would yield a reduction in the coefficient on co-posting of 0.65, or 2/3rds of the unconditional effect. Similarly, for momentum, the coefficient of 0.097 suggests that stocks in the opposite ends of the spectrum of return momentum would have a reduction of coefficient of 90%. The next two columns report results for value stocks, the loading on the book-to-market decile is an insignificant -0.003, but the coefficient on β_{HML} decile is a statistically significant -0.039. One potential reason is that book-value, as commonly defined in the academic literature, tends to lose a great number of observations, particularly for growth stocks. Growth stocks might be the most affected by social media. To the extent β_{HML} is correlated with value (albeit a covariance itself), it suggests that a similarity in exposure to value amplifies the relation between social media activity and return correlation. Next, investment-to-assets and β_{CMA} both support the idea that investors view investment aggressiveness as a style class. Stocks co-move more within beta and dividend deciles, as predicted. Also, in unreported tabulations, I find that interaction terms on characteristic differences in spreads, ownership and analyst coverage do not load negatively, which is sensible as I conjecture these are not style classes along which investors sort stocks.

²⁶Economically, “category” investing suggests that investors place stocks into rankings or styles and would anchor on relative ranking. . Further, the cross-sectional dispersion of the characteristic may vary over time. I try to immunize my results from changes in this distribution. I try a few specification checks, which yield consistent results. Instead of defining the absolute difference in decile, I define a boolean as being at least 3 categories apart. This yields similar and in fact somewhat stronger inferences depending on the coefficient.

Table 1.11: Sentiment

In this table, I explore whether investor-reported sentiment affects the relation between social media activity and comovement. The outcome variable is excess return correlations at the monthly horizon, from 2012-March2016. Of interest is the interaction of $\log(1 + Coposts_{ijt})$ and a sentiment measure. For the calculation of $netSent$, I assign $\$Bullish$ cashtags a 1, $\$Bearish$ a negative -1, and either equal weight or weight by followership as indicated. The variable $1\{Disagree\}$ is a measure evaluates to 1 if the net-sentiment throughout the month for stocks i and j are disagreeing in sign (one positive, the other negative). $NetSent$ is defined the signed log count of the absolute sentiment. For example, net -5 posts is $\log(1+5)*-1$. The triple interaction term is such that when both $netsent_i$ and $netsent_j$ are positive, they are in agreement. Finally, $\frac{Sent_{it}}{Tags_{it}}$ is a ratio of net sentiment to the number of sentiment-related posts. Standard standard errors are double-clustered by pair-month, with pair-month fixed effects. For brevity, unconditional sentiment is suppressed.

| | | | | | | | |
|--|---------------------|----------------------|---------------------|---------------------|---------------------|----------------------|----------------------|
| $\log(1+Coposts_{ijt})$ | 0.760*** (0.056) | 0.761*** (0.057) | 0.759*** (0.057) | 0.767*** (0.059) | 0.765*** (0.062) | 0.831*** (0.057) | 0.807*** (0.056) |
| $\log(1+Coposts_{ijt}) * 1\{disagree\}$ | -0.276** (0.126) | -0.294*** (0.111) | -0.276** (0.130) | | | | |
| $\log(1+Coposts_{ijt}) * netsent_i$ | | | | -0.010 (0.007) | -0.008 (0.025) | | |
| $\log(1+Coposts_{ijt}) * netsent_j$ | | | | -0.007 (0.009) | -0.033 (0.033) | | |
| $\log(1+Coposts_{ijt}) * netsent_i * netsent_j$ | | | | 0.003* (0.002) | -0.016 (0.021) | | |
| $\log(1 + Coposts_{ijt}) * \left \frac{netSent_{it}}{tags_{it}} - \frac{netSent_{jt}}{tags_{jt}} \right $ | | | | | | -0.175*** (0.040) | -0.135*** (0.043) |
| FE? | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Follower-Weighted Sentiment? | No | Log | No | Log | No | Log | No |
| Num. obs. | 7,061,890 | 7,061,890 | 7,061,890 | 7,061,890 | 7,061,890 | 7,061,890 | 7,061,890 |
| Adj. R ² | 0.05065 | 0.05065 | 0.05065 | 0.05065 | 0.05065 | 0.05065 | 0.05065 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

1.7.3 Investor-reported sentiment

Next, I interact co-posting measures with sentiment, as defined by social media posters' use of cashtags (e.g. *\$Bearish* or *\$Bullish*). Sentiment is not randomly assigned, but it would be puzzling if the co-posting-correlation relation does not abate, or instead amplifies, when the *observed* attitudes of investors about these two stocks diverges. We may expect reduced buying or even net selling by social media consumers when pessimism occurs (although there will be gross-buying due to disagreement or contrarian views). In general, these cases are somewhat rare; StockTwits, at least, is generally a conduit for expressing *\$Bullish* sentiments.

For monthly results, I report statistics based on contemporaneous sentiment reported in the same month. For weekly, I use the sentiment calculated from the contemporaneous and prior week sentiment. The reason is that relative to posters covering a specific topic by tagging the stock, less 20% of these posts (7.5 million of 37 million) report a sentiment. Also, it is sensible that a trader on StockTwits would take into account recent sentiment and not only the currently reported sentiment.

I specify sentiment disagreement in many ways. I first describe it as a disagreement in sign - when the net sentiment across StockTwits is positive on one stock and negative on another. Second, I define sentiment as a triple interaction of the logged net sentiment on the stock, which is an intensity or scale measure of disagreement. If there are five bearish tags, and 10 bullish tags, the net sentiment is $\log(5+1)$; the converse is $-1*\log(5+1)$. Third, I define a difference in net sentiment ratios to the number of overall sentiment posts. To calculate a ratio, I take the net-sentiment divided by the total number of posts conveying a sentiment. Four positive posts out of five posts conveying a sentiment would yield a ratio of .6, based on the numbers $(4-1)/5$. The absolute difference of these ratios is a measure of disagreement. For these posts, sentiment is sometimes follower weighted, equal weighted, or log follower weighted.

Table 1.11 reports the monthly results. I report only the monthly results for brevity. Seven of eight specifications support the hypothesis that disagreement abates the posting-correlation relationship. Columns 1 through 3 reliably show that co-posting is one-third less effective when the net

sentiment on the two stocks are both non-zero and pointing in opposite directions. In column 4, the triple interaction term suggests that when the *netSentiment* is in the same direction, the co-posting sentiment relation is amplified; conversely, when they disagree, the co-posting relationship abates. However, the effect is economically small; the standard deviation on the *netSentiment* measure in Column 4 is about 2.4, suggesting the standardized coefficient is about 5.5 times larger or 1.54 basis points. Column 5 repeats the same but is insignificant, possibly because without taking into account the followership of the posters, the reported sentiment has no influence. Column 6 and 7 shows the impact on the co-posting relationship defining disagreement as the difference of sentiment ratios. The coefficients imply that maximal disagreement (fully bearish investors on one stock, and fully bullish investors on another) would reduce the co-posting relationship by somewhere between 27-35 basis points of a roughly 82 basis point effect, similar in magnitude to the estimate of the first three columns. Overall, at the monthly level, the net effect of posting and correlation is not negative when overall sentiment is negative. Perhaps disagreement combined with short-sale aversion may together to partially limit the clout of pessimistic investors. Specifically, partially negative sentiment on one stock may result in reduced buying pressure as fewer people sell than buy, but most pessimistic investors sit out if they start out with zero inventory.

I relegate the weekly analysis to the Appendix. The overall inference carries over. The posting-correlation relationship is in fact far more economically significant, such that the effect of postings on correlation is actually completely abated or net negative. However, the results are less reliable, significant in half of specifications.

1.7.4 Other types of comovement

The main focus of this chapter has been return correlation, but in this section I repeat the main result using other types of correlation: absolute returns, trading activity and liquidity. The main goal is to augment narrative that social media causes return comovement - if social media activity relating two stocks increases return correlation, one might this to be implemented by trading activity occurring at the same time. I also explore the implications for absolute returns and liq-

Table 1.12: Other types of comovement

In this table, I repeat the main inference in Table 4 with correlations of alternative outcome variables. At the monthly level, all values are based on CRSP. The Spread is the CRSP *ASK* field minus the CRSP *BID* field. The percent spread is based on the close price. The Amihud value defaults to 0 (infinite inelasticity) when no trading occurs, which is a very small fraction of trading days. At the weekly level, all values are computed from half-hour bars from the Daily Trades and Quotes National Best Bid or Offer or the Trades File. The spread is based on the half-hour end quote, the percent spread is calculated as a percentage of the midpoint. I report robust standard errors, double-clustered by pair-time, with pair-time fixed effects.

| Panel A: Monthly | | | | | |
|--------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| | ρ of ... | | | | |
| | $ \text{ret}^e $ | Turnover | Volume | Amihud | % Spread |
| $\log(1+\text{Coposts}_{ijt})$ | 3.619*** (0.218) | 5.307*** (0.372) | 5.513*** (0.373) | 0.590*** (0.118) | 0.316*** (0.085) |
| $\log(1+\text{Posts}_{it})$ | -0.217*** (0.039) | -2.633*** (0.196) | -2.569*** (0.188) | -0.494*** (0.090) | -0.385*** (0.067) |
| $\log(1+\text{Posts}_{jt})$ | -0.255*** (0.036) | -2.711*** (0.210) | -2.634*** (0.198) | -0.452*** (0.102) | -0.328*** (0.062) |
| Controls? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| FE? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| Num. obs. | 7,059,513 | 7,061,401 | 7,061,401 | 7,061,401 | 7,061,401 |
| Adj. R ² | 0.017 | 0.140 | 0.138 | 0.089 | 0.036 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

uidity. Under heterogeneous beliefs, comovement in absolute return correlation could conceivably see sharper increases than comovement in returns - the impact of buy order flow is attenuated by disagreeing investors who sell against it. We may also expect liquidity comovement as social media consumers trade heavily, taking liquidity and causing pressure in market variables. I measure market activity by turnover and volume. I measure liquidity using the Amihud illiquidity measure, percent spread, and order imbalance (in the case of weekly results). For this analysis, I report both weekly and monthly results and their relation to co-posting activity. In Appendix A.5, I report the summary statistics for all the correlation variables I use as dependent variables in Table 1.12.

Table 1.12 reports the results. Panel A reports monthly results and Panel B reports weekly results. Turning to Panel A, column 1 reports a coefficient on absolute return comovement of 3.64. The coefficient is four times as large as the estimate for excess return correlation, although the

Panel B: Weekly (robustness)

| | ρ of ... | | | | | |
|--------------------------------|-------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| | $ \text{ret}^\epsilon $ | $ \text{OIB} $ | Volume | Amihud | OIB | % Spread |
| $\log(1+\text{Coposts}_{ijt})$ | 1.346*** (0.076) | 0.857*** (0.100) | 2.493*** (0.175) | 0.256*** (0.048) | 0.072 (0.050) | -0.289*** (0.089) |
| $\log(1+\text{Posts}_{it})$ | -0.118*** (0.031) | -0.226*** (0.066) | -2.150*** (0.068) | -0.194*** (0.039) | -0.089*** (0.022) | -1.820*** (0.069) |
| $\log(1+\text{Posts}_{jt})$ | -0.089*** (0.029) | -0.220*** (0.070) | -2.070*** (0.070) | -0.190*** (0.039) | -0.090*** (0.019) | -1.790*** (0.061) |
| Controls? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| FE? | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> | <i>Yes</i> |
| Num. obs. | 42,933,800 | 42,733,126 | 42,733,126 | 42,349,034 | 42,733,126 | 42,732,462 |
| Adj. R ² | 0.13 | 0.29 | 0.29 | 0.09 | 0.01 | 0.41 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

mean absolute return comovement more than two times as large. The next two volumes report turnover and volume. The coefficients on co-posting of 5.3 and 5.51 are seven times as large as the coefficient for raw return, but the average correlation of turnover and volume, unconditionally, is 13.4 and 12.8 respectively, or about 12 times larger. Thus, in overall economic magnitudes, the effect is actually somewhat smaller than the effect on returns. The final columns 4 and 5 reports estimates for commonality in liquidity. The coefficients of 0.59 and 0.316 are positive and precisely estimated. Taken together, it seems that social media activity also is related to commonalities in liquidity and trading activity at the monthly level.

Panel B reports the weekly results. The inferences are similar, except the inference for liquidity is mixed. It seems that social media activity explains commonalities in Amihud illiquidity, it does not explain commonalities in spreads. Also, the inference for order imbalance is only weakly positive. Regardless, the evidence strongly points toward increased commonality in market trading during periods of social media activity, suggesting that perhaps the mechanism for social media impacting stock return correlation is through synchronized trading.

1.8 Conclusion

In this chapter, I show that social media activity that associates two stocks increases within-pair, within-time excess return correlation. The baseline correlations are not obviously related to corporate news events or industry events. To argue causality, I show that concurrent postings that reach pre-determined larger audiences tend to comove more, both at the weekly and monthly level. In addition, using a subset of the sample for which I continuously captured the StockTwits Trending banner at a high frequency, I provide some evidence that social media guides poster discussion and increases the tendency of posters to “co-post” about stocks. I then provide moderately reliable evidence showing that instrumented co-postings also increase excess correlation. Finally, I show that many logical corollaries of social media consumers causing comovement are consistent with the evidence: that posters with larger followings impact markets more, that stocks with greater susceptibility to small trader influence, or which exist in a common investing category, comove more during social media episodes, and that sign disagreement in the sentiment of stocks abates the observed relation between social media activity and correlation.

The magnitudes in this chapter are economically large but plausible. Israelsen (2014) reports that a standard deviation increase in analyst common coverage increases annual comovement (of daily returns) by about 3% - my causal estimates are about a quarter to a half in size, using a much more conservative estimation methodology. Also, it is worth noting that StockTwits is a market leader, but at most a large plurality in financial social media. Separate discussions held on Twitter, Seeking Alpha, Estimote, or the smaller sites hoping to make a mark in this exciting space, are likely dynamic and distinct from the conversations held on StockTwits.²⁷ This suggests the findings of this chapter are likely understated, especially as social media continues to grow.

It will be interesting to see future research utilize the fascinating institutional details of social media to identify other causal effects. For example, it is possible social media activity changes

²⁷As far as I know, since 2011, none of the sites are perfectly integrated, allowing posters to cross-pollinate their content. Originally, Twitter allowed StockTwits posters to mirror their posts to both. However, they removed this integration as StockTwits grew. When Twitter acquired Gnip, an application programming interface for serving social data, StockTwits was removed from the menu of offered services. It is clear the two compete.

the disclosure policy of firms who may want to capitalize on increased liquidity or, alternatively, strategically avoid high-attention periods. Another possibility is that social media sentiment may feed back into the real decisions of firms, particularly for consumer retail firms. And, of course, as social media platforms begin to aggregate and classify non-textual information, social media can increasingly advance as a window into investor beliefs and behavior. For example, some platforms crowdsource detailed analysis of firm filings. Finally, with my current empirical design, I cannot determine whether the influence of social media reflects a transmission of information or noise, and as a result, I cannot make a welfare statement. However, I plan to explore this in the near future.

Chapter 2

Who Should Regulate Investment Advisors?

2.1 Introduction

The registered investment advisory industry is a primary conduit through which investors access financial markets. In 2015, the industry oversaw over \$66 trillion in assets. The Survey of Consumer Finances suggests that 64% of all American households are a customer of some RIA. When working for advisory firms, their representatives uphold a legally enforceable fiduciary duty to act in a client's best interest, as established under the Investment Advisers Act of 1940. However, after giving advice, adviser representatives often are asked to implement their recommendations, acting as *brokers* who profit through commission. Wearing these two hats creates conflicts of interest widely thought to subtract from the quality of financial service, and this incentive problem is replete throughout the industry.²⁸

Perhaps in large part to this incentive problem, prior research headlined by Egan et al. (2016) document that financial advisors frequently engage in misconduct. Misconduct has direct financial consequences on victims and also lowers general trust in the financial sector, possibly deterring marginal savers from building wealth. One might assume that, over time, self-correcting market forces would lead to extermination of misbehavior. However, Egan et al. (2016) argues that markets appear to tolerate misconduct, observing high recidivism and an ease of finding employment for even those terminated for cause. That markets fail to fully self-regulate motivates a need to understand what factors impact the quality of internal or external governance. An important lever for governance of financial institutions is regulatory design. In this chapter, we ask whether the scope of a regulator - whether a local or national regulator - better deters advisor misconduct.

We exploit a unique regulatory shift due to the Dodd-Frank Act to study how regulatory jurisdiction affects advisor misconduct. Assessing the importance of regulatory jurisdiction is difficult

²⁸In 2012, various SEC documents suggest that 87% of investment adviser representatives (individuals working for an investment advisory firm) are also registered in the capacity as a broker-dealer. In 2015, this number has climbed well past 90%.

as it rarely changes, and when it does, laws tend to change as well. Our setting bypasses these standard challenges. The Dodd-Frank Act caused the SEC to transfer oversight of “mid-sized” investment advisers (\$25-100 million in assets under management) from the SEC to state regulators, except for advisers located in Wyoming and New York.²⁹ This decision was announced on July 21, 2011 and in effect by January 1, 2012. The impetus for the shift was exogenous to mid-size adviser behavior. The intention was to free up SEC resources so that it could increase oversight of hedge funds and private equity firms. The size threshold was chosen out of convenience, as it would reverse a component of the 1996 National Securities Market Investment Act (NSMIA). NSMIA had assigned mid-size advisers to the SEC as part of a broader effort to unify state-securities regulations. Just over 38% of all existing SEC-registered firms were affected by this re-jurisdiction.

Using a differences-in-differences design, we study how a shift from SEC to state-regulator oversight affects the probability of a customer complaint. Complaints are a good measure of adviser misbehavior and better than available alternatives. Complaints are publicly disclosed and observable for every professional in the industry. In addition, complaints are preference-adjusted; regardless of preferences, complaints reveal the perception of substandard adviser advice. The most obvious alternative, studying investment returns, is sparsely available, and would yield unsuitable comparisons across clients of different preferences. Studying regulatory sanctions may instead represent change in regulator behavior, not adviser behavior. We construct a survivorship-bias-free panel dataset at the representative-year level. We narrow the time period to the three years before and after the implementation of Dodd-Frank (2009-2014). We assemble this data using a variety of regulatory disclosures made by the SEC.

Our main finding is that complaint rates increased by 50% relative to the unconditional probability of 0.8 percentage-points per year. The result is robust to a variety of sample splits: only advisers around the \$100M threshold, across-state (using only mid-sized advisers in Wyoming and New York as the control group), within-state (excluding Wyoming and New York), and in matched samples comparing representatives with similar complaint histories. Whatever baseline we choose,

²⁹Wyoming had no registration requirements for investment advisers at the time and New York does not examine investment advisers.

the result is the same: client complaint rates increased for newly state-registered advisers after Dodd-Frank. Visual diagnostics suggest the increase persists through 2014.

Examining the details of individual complaints, the increase in complaints seem to be advisory-related, alleging issues such as fraud, misrepresentation, suitability, and unauthorized activity. We find no statistically significant response in “broker-related” categories of complaints, such as fees and churning. This is sensible because advisory-related activities are under state or SEC purview, which changed as a result of Dodd-Frank, whereas broker activity is FINRA-regulated, whose jurisdiction did not change in the same way. Also, the increased incidence of complaints did not represent frivolous charges. If anything, alleged damages rose by about \$100,000 per complaint with no significant change in outcome probabilities for the re-jurisdictioned advisers, relative to the SEC-registered advisers.

Our next analysis aims to disentangle two explanations for the observed increase in complaint incidence. First, actual misconduct may have increased, which results naturally in more filed complaints. Second, complaints may be more frequent because state regulators somehow elicit more complaints. State regulators may be more attentive to mid-size firms, given mid-size firms are the largest at the state level but the smallest at the national level. Local regulators might have also better soft information. To distinguish these explanations, we turn to the cross-section. If a firm is more well-regulated, and receives more complaints, this is consistent with a monitoring story. Weaker regulation coinciding with more complaints would be consistent with our main result reflecting increased misconduct.

A variety of cross-sectional supports increased misconduct, not better monitoring. First, if the increase in complaints is due to better monitoring, we would expect the treatment response to be stronger in better-equipped states with high regulator-staff-to-adviser ratios. We find *instead* that the treatment effect is highest in states with the lowest ex-ante staff-to-firm ratios. Second, if local regulators exert effort or capitalize on soft information, we expect then the treatment effect to be highest for firms physically near by the state regulator. Instead, we find that a standard deviation increase in distance to the regulator resulted in a 50% greater treatment response. This

finding is robust to a variety of fixed effects that account for how distance may be correlated to local, time-varying economic conditions around where the firm is headquartered. Third, if regulatory dilution occurred, then the always state-registered advisers (sub \$25 million in AUM) should also respond to treatment. Indeed, we find that relative to always-SEC-registered advisers, state-registered advisers also increase their complaint rates, although by a lower amount than the transitioning firms.

Another possibility is that regulators were not more proactive, but customers view them as sympathetic to or efficient with customer complaints. That is, a change of regulatory scope may elicit strategic customer responses, as opposed to adviser responses. We present two tests. First, if the increased complaint incidence represents forum-shopping, complaints should rise the most among sophisticated clients. Instead, we find that complaints rose most for adviser representatives working in counties with less educated people, who also tend to be older. These findings are echoed at the firm-level, where we find treatment effects were lower among firms with more sophisticated clients. Finally, if somehow regulator or client awareness increases, the equilibrium response by advisers with poor track records should be reduced recidivism, knowing they are at-risk to strategic complaints. Instead, we find that representatives with more past complaints saw the greatest increases in customer complaints. Our six cross-sectional tests rule out the two obvious alternatives to a misconduct story.

Lastly, we aim to shed light on at least one possible mechanism for our findings. Some, such as reputation or competence, are difficult to assess in the data. State budget constraints seem partly responsible for the higher misconduct. Using hand collected data, we find that on average the budget requested by state securities departments did not respond despite significant increases in oversight responsibilities. This fact alone works against an increased monitoring explanation. While many states regulators explicitly mentioned additional oversight workloads due to Dodd-Frank in their 2012 budget filings, most states faced budget crises during this period due to the aftermath of the financial crisis. Consistent with financial constraints driving regulator effectiveness, we find that even existing state-registered advisers saw an increase in complaints, albeit smaller than the in-

creases for those that switched from the SEC. If state monitoring ability increased or remained the same, then misconduct by previously state-registered advisers should decrease or stay the same.

We make two contributions. First, we contribute to the literature on the optimal design of financial regulation. The paper most relevant to ours is Agarwal et al. (2014), who find that jurisdictional changes between state and national banking regulators appears to create inconsistent implementations of banking regulation.³⁰ Our work asserts that national regulators may be stronger, and suggests at least one mechanism through which local regulators may underperform super-local ones: a lack of local regulatory resources. This speaks more broadly to potential coordination failures between national and state regulatory agencies. However, other mechanisms, such as institutional human capital or reputation, are left untested by our work.

The other area we contribute to is the nascent literature in the investment adviser industry. Dimmock et al. (2015) study the peer effects of investment adviser fraud, after advisory firms merge. Foerster et al. (2014) study the quality of investment adviser recommendations, finding they outweigh conflicts of interest in terms of costs to client net worth. Gurun et al. (2015) study the role of trust on use of investment advisers, using the Madoff scandal as a natural experiment. Egan et al. (2016) document widespread misconduct and recidivism in the investment adviser industry and suggest that market incentives and firm governance alone do not seem sufficient in eliminating misconduct, motivating a role for the regulator to investigate and publicize information to clients. Finally, our results are also of significant interest to recent policy discussions re-evaluating Dodd-Frank and debates on holding brokers to the controversial Department of Labor fiduciary standard.

2.2 Oversight of Investment Advisers and Broker-Dealers

2.2.1 History of Regulatory Jurisdiction

In the United States, investment advisers are regulated under the Investment Advisers Act of 1940. The legal definition of an investment adviser is broad: an adviser is “any person who, for

³⁰Although unlike in Agarwal et al. (2014), regulatory arbitrage is less relevant in our context, as the impact of our regulatory change is meant to be permanent, and the low lead time to implementation reduces concerns about partial anticipation.

compensation, engages in the business of advising others, either directly or through publications or writings, as to the value of securities or as to the advisability of investing in, purchasing, or selling securities, or, who for compensation and as part of a regular business, issues or promulgates analyses or reports concerning securities”, except when “solely incidental”.³¹ The individuals employed by investment adviser firms are termed investment adviser representatives. The Advisers Act holds investment advisers to a fiduciary standard, requiring investment advisers to act in their client’s best interests. Advisers are regulated by the SEC or state regulator depending on the adviser’s assets under management.

In addition to providing securities advice, investment advisers may manage investment portfolios, provide financial advice, and offer brokerage services (such as buying or selling stock or bonds).³² Usually, an investment adviser firm has several investment companies (mutual funds, closed-end funds, unit investment trusts, private funds) and in turn each investment company could offer several different funds. Thus, common names for investment adviser representatives include asset managers, investment counselors, investment managers, portfolio managers, and wealth managers.

Almost 90% of investment advisers and representatives are also registered as broker-dealers as of 2012. Brokers are not held to a fiduciary standard but rather to a “suitability” standard of conduct. When a broker recommends buying or selling a security, the broker must consider a client’s income and net worth, investment objectives, risk tolerance, and other security holdings. Differences in compensation contracts between brokers and investment advisers lead to conflicts of interest. Brokers typically receive commissions and product fees. Investment advisers typically earn a fee based on assets under management. Clients who do not understand when their adviser

³¹“Solely incidental” in laymen terms is meant to construe those whose course of business contains content that could be construed as investment advice, but not for the purpose of giving such advice. Business school professors in finance, for instance, do not have to register as investment advisers if discussing efficient market theory or teaching CAPM.

³²

Six years prior to the Investment Advisers Act of 1940, Congress wrote into law the Securities Exchange Act of 1934, which defined a broker-dealer as “any person or company engaged in the business of buying and selling securities on behalf of its customers, for its own account (as dealer) or both.”

is acting as a broker or investment adviser may misinterpret the incentives driving the advice. The overlap in roles also creates an overlap in regulatory jurisdiction. Brokers are regulated by the Financial Industry Regulator Authority (FINRA), a self-regulating organization. The overlap is so deep that the SEC actually depends on FINRA to maintain its data on investment advisers.

Responsibility for overseeing investment advisers has shifted between state and federal governments multiple times over the past several decades. Prior to 1992, the SEC and state regulators had concurrent regulatory authority over the investment adviser industry. The National Securities Markets Improvement Act of 1996 (NSMIA) folded mid-size advisers into SEC jurisdiction, as part of a collection of efforts to integrate national securities regulations and increase the power of the relevant national regulator. Since NSMIA circumscribed their authority, state securities regulators have been aspiring to reclaim it. As the North American Securities Administrators Association (NASAA) Executive Director Russ Iuculano stated, *“The financial catastrophe of 2008 gave NASAA a great opportunity to make its case that our system of financial services regulation must be strengthened, but only through the combined efforts of state and federal regulators.”*

On July 21, 2010, on the heels of the Great Recession, President Obama signed into law the Dodd-Frank Wall Street Reform and Consumer Protection Act. Section 410 of the Dodd-Frank Act raised the AUM threshold for state regulation of investment advisers from \$25 million to \$100 million. The primary motivation for this change was that Title IV of the Dodd-Frank Act repealed the “private adviser” exemption, which had exempted hedge funds, private equity firms and venture capital firms from registering with the SEC. Raising the AUM threshold freed up SEC resources to monitor this new cohort, except where state regulators did not subject advisers to regular examination (Wyoming, and New York, and originally Minnesota until recent regulation). Figure 2.1, acquired from the SEC announcement, depicts the timeline of the events.

As of the time of the law’s announcement, the SEC estimated 3,200 investment advisers would be delegated to examination by state law. In particular, 3,512 investment advisers had, as of 2011, filed with between \$25-90MM in regulatory assets under management, and about 300 would be exempt on the basis they are either foreign advisers, or did their business principally in states where

investment advisers are not regulated by a state agency delegated the responsibility of security oversight.

Financial professionals registered as investment adviser representatives or broker-dealers are generally required to maintain updates with regulators of any material event that clients or employers may find relevant. Beyond regulatory events and customer complaints, required updates include disclosures of personal bankruptcy, civil suits, or liens on their personal assets.

Our workhorse variable will be the number of complaints initiated by customers, regardless of their status.³³ These include complaints that are in progress, settled, denied, or withdrawn. This is different than the definition of misconduct used in Egan et al. (2016), who consider the category, “Employment Separation After Allegations” to be part of misconduct, and ignore the category “Customer Dispute - Denied”, as well as other customer disputes. Their purpose is to identify misconduct of any kind, while ours is to identify misconduct specifically from the perspective of customers. To the extent our consideration of denied complaints could drive our results, we analyze the propensity for a complaint to be denied in Section 5.

How costly is it to file a complaint? Figure 2.3 examples the electronic procedure for filing complaints for the state of New Jersey. Methods for other states are nearly identical. There is no fee to file a complaint, and complaints are filed electronically. Of course, clients do incur costs in other ways - carrying through with complaints takes time, can lead to alienating a relationship with an existing adviser, and reclaiming damages through arbitration or settlement requires processing fees, as well as potential legal costs.

The IAPD stores all complaints with alleged damages greater than \$5,000 filed in the past 10 years. Figure 2.4 examples some complaints. All complaints contain the date received and a current status (e.g., pending, settled, denied). Often times, the complaints will contain a product code. For example, 20% of complaints dealt with stocks and 4% with over-the-counter securities. The complaint data contain unstructured text detailing the nature of the allegation from the different participants (e.g., regulator, adviser, client), as well as docket identifiers to track legal proceedings

³³Customer complaints in BrokerCheck may undergo a variety of status updates. Complaints that are executed fully are often arbitrated through FINRA’s arbitration process or processed by some other formal procedure.

| July 21, 2011 to December 31, 2011 | January 1, 2012 | March 30, 2012 | June 28, 2012 |
|---|---|---|---|
| New registration thresholds and requirements apply to new applicants, but not to existing SEC-registered advisers until the dates indicated in this table, as applicable. | Each SEC-registered adviser as of July 21, 2011 must remain registered with the SEC until this date (unless relying on an exemption). | Last day for all SEC-registered advisers to file the required Form ADV amendment. | Mid-sized advisers not eligible for SEC registration must file form ADV-W to withdraw by this date. |

Figure 2.1: Event Timeline

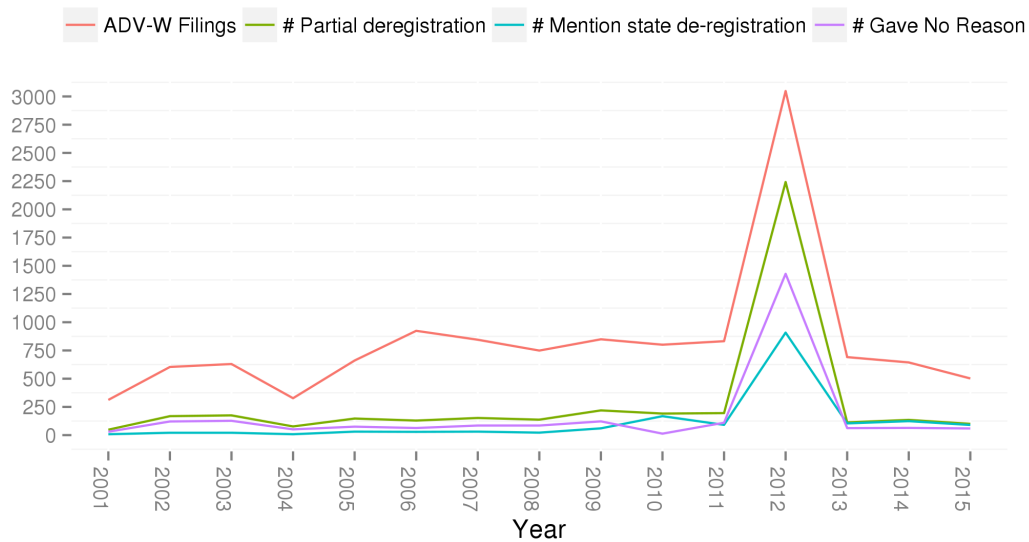


Figure 2.2: Annual ADV Deregistration Filings

This graph shows that investment adviser firms de-registered from the SEC in 2012 in response to the Dodd-Frank Act. To indicate a change in registration status (cessation, merger, or partial deregistration), firms file a Form ADV-W. In the graph, the line “ADV-W Filings” refers to the total number of filed Form-ADV-Ws. “Partial de-registration” presents the number of Form ADV-Ws for a partial deregistration. An optional field in Form ADV-W allows firms to specify the reason for partial deregistration. The line (“# Mention state de-registration”) indicates how many partial de-registrations specifically mentioned the intention to register with state-securities regulator.



File a Complaint

The Bureau of Securities investigates complaints against individuals and firms selling securities or offering investment advice as well as companies issuing securities investments. The Bureau is empowered to bring administrative actions or civil law suits to enforce the registration and anti-fraud provisions of the New Jersey Uniform Securities Act. The Bureau may refer certain matters for criminal prosecution.

Please be advised that the Bureau does not have the specific authority to order restitution or the repayment of any monies which you may believe are due you.

Investor Information

Name:

Street Address:

City: State: ZIP Code:

Daytime Number: Evening Number: Fax:

Email Address:

Firm Information

Firm Name:

Street Address:

City: State: ZIP Code:

Telephone Number (1): Telephone Number (2):

Email Address:

Complaint Information

1. Type of firm (if known):
If other, please specify:

2. Name and title of firm's agents or employees with whom you dealt:

Name:

Title:

Figure 2.3: Filing a Customer Complaint

Figure illustrates how customers file a complaint with the New Jersey Securities Regulator.

| Individual CRD | Date Received | Date Addressed | Status | Product | Alleged Damages | Settlement Amount |
|---------------------------|----------------------|-----------------------|---------------|------------------------|----------------------------|------------------------------|
| 1006724 | June 20, 1994 | April 1, 1995 | Settled | Other | \$316,000.00 | \$25,000.00 |
| 1006724 | January 24, 2002 | January 29, 2002 | Arbitration | Equity - OTC | \$3,000,000.00 | |
| 1001126 | November 22, 2002 | January 10, 2003 | Denied | Other | \$29,884.00 | |
| 1001136 | October 10, 2003 | January 23, 2004 | Settled | Unit Investment Trust | \$20,000.00 | \$20,627.00 |
| 1000034 | January 24, 2006 | March 7, 2006 | Denied | Annuity | \$28,178.00 | |
| 1007710 | August 3, 2009 | December 2, 2010 | Settled | Lehman Special Purpose | \$0.00 | \$240,000.00 |

Figure 2.4: Complaint Example & Summary Statistics

The figure shows some examples of customer complaints of varying degree. Disputes can be made against the investment adviser firm, their representatives, or both. Disputes are automatically stored and appear on FINRA's BrokerCheck and the IAPD if the alleged damages are over \$5,000, or if the dispute resulted in some legal action, regardless of legal outcome.

related to a case.

2.3 Data & Methodology

2.3.1 Data collection

Data for this study comes from Form ADV filings filed via the Investment Adviser Registration Depository (IAPD) with the appropriate SEC or state-securities regulator. Using the Freedom of Information Act, we obtained the universe of Form ADV filings from 2000-2015. We identify firms that switched from SEC oversight to state oversight in 2012 using Form ADV-W, which indicates a partial deregistration with the SEC.

We also retrieve the disclosure histories of investment adviser representatives from the IAPD and FINRA's BrokerCheck databases. Both filings to the SEC and FINRA are stored in the Central Repository Deposit (CRD) database. The close relation between the SEC and FINRA also allowed linking SEC Filings to FINRA BrokerCheck data through unique CRD identifiers. Both individuals and firms are given unique CRD numbers. The difference between the two sources is that about 70,000 investment adviser representatives never registered as brokers and are thus not in FINRA's

database.³⁴ Because there may also be reporting differences between the two sources, we merge complaint information from both the IAPD and FINRA BrokerCheck. To obtain detailed complaint information, we use a web-scraper that takes a representative's CRD number and queries the IAPD website. We query all CRD numbers between 1 and 10 million. We then download the detailed reports in PDF form and extract text using open-source software Tabula and Poppler.

2.3.2 Sample Construction

We develop a data set where for each person-year we have the employing firm, complaints received, and relevant individual characteristics. We match representatives to firms using each firm's unique CRD number, an internal identifier used by NASD/FINRA and securities regulators. In each representative's detailed reports from the IAPD, we see the representative's full employment history at the branch level, including start and end dates.³⁵ The IAPD tracks data for any person employed within the last 10 years.

The resulting sample is an annual panel consisting of individual CRDs and firm CRDs, and the branch location worked at while the individual was there. We also observe other individual characteristics, such as professional designations and exams passed (Series 66, Series 63, Series 65). The final data we extract is on disclosures. These include customer complaints, criminal actions, regulatory actions, litigation, terminations, civil suits and other financial matters such as liens that might be pertinent to an adviser's ability to manage money. We extract all fields and normalize across complaint types. The two difficulties are that the field names and data are hand-entered and therefore must be cleaned, and that the different disclosure types have different names for what is effectively the same field (date litigation filed, date complaint received). After this

³⁴On the other hand, approximately 700,000 brokers do not register as investment advisers representatives.

³⁵This section is apparently maintained by hand, as sometimes branch locations are misspelled (ATALANTA, ATLANTA, representing ATLANTA, for example). We normalize the names and remain conservative in lumping branches together. For our analyses, this is a conservative approach because an extra branch fixed effect would simply chip away from our point estimates of interest. The second unrelated data issue is that in less than 0.25% of individual years, advisers belong to more than two firms (one they are leaving, one they are going to), owing evidently to cases where the firm has multiple CRD numbers assigned. These cases are rare in the data. In these cases, we defer to the first method, which assigns individuals to one firm at a point in time, or assign the individual to the firm CRD most commonly seen in the sample.

process, we extract the disclosure and assign a date to it. Similar to prior literature, we assign a complaint to the year in which it was officially received by the CRD. If an individual did not receive a complaint, they are assigned a 0 for that year. We drop about 7% of disclosures that do not include the date received.

2.3.3 Methodology

We hypothesize that a regulator’s monitoring capability acts as a deterrent to misconduct. The ability to monitor depends on the resources of the regulator. One might assume that the SEC, as the incumbent regulator for larger firms, attracts higher-human-capital staff and has more experience auditing more complex and larger firms. In contrast, state regulators may have a local information advantage. However, having not monitored mid-size advisers for 15 years, state regulators had a lot to learn about their local mid-size advisers. Our empirical study identifies the average net impact of these potential effects.

The empirical specification takes the form:

$$Complaint_{it} = \alpha + \beta_1 Treated + \beta_2 Post^{2012} + \beta_3 Post^{2012} \times Treated + \varepsilon_{it}$$

where the coefficient of interest is β_3 . We perform difference-in-difference analyses using two methods. The first is an annual panel, and the second is a collapsed-three-year window approach in the manner of Bertrand et al. (2004). We pick a three-year window around 2012 because complaints may not be filed immediately when the misbehavior happens. Thus, examining complaints filed during the three-year period gives us more power.³⁶ Clients have 6 years from misdeed to file. The trade-off between the two methods is that the collapsed window approach has the advantage that serial correlation of residuals do not mislead inferences, while the annual approach mitigates the issue of parallel trends or capturing some other event. Also, since individuals may move across

³⁶ Shortening the window to 2 years results in qualitatively similar inferences.

firms, annual panel estimates mitigate this issue.

The initial outcome variable is the probability of receiving a complaint - the extensive margin incidence of complaints. We also examine the intensive margin by examining the Act's effect on (1) the log count of complaints, or (2) the log(amount of alleged damages). However, not all complaints seek alleged damages (or the field is not properly populated by the record-keeping agent). Our outcome variable differs from Egan et al. (2016) in that we look for complaints, regardless of outcome, whereas they look at complaints that resulted in sanctions and terminations. Our measure is different because terminations may be initiated by the regulator or the firm, and confound our interpretation.

Our main specification will be within-firm, within-period. In the annual panel, this means including year fixed effects. We cluster standard errors at the state level as each state may have implemented the regulation differently and residuals may be correlated within a state, but are likely independent across states. We also clustered at the firm level, and sometimes results improved slightly and sometimes results worsened slightly.

2.3.4 Identifying Treatment and Control Groups

We label firms as “Treated” if their 2011 assets under management are below \$100 million, they file a Form ADV-W indicating partial de-registration (full de-registration implies a business cessation or change of ownership), and they are located in the affected states (all but New York and Wyoming). Our criteria likely capture the vast majority of firms actually affected by the re-jurisdiction. When discussing the regulation change in early 2011, the SEC projected around 3,200 firms would be affected. We find 2,316 firms. This smaller treatment sample is likely due to the bull market in equities that increased adviser's AUM. Consistent with this story, an article discussing the event suggests the SEC reports “over 2,300” firms made the switch through October 2012³⁷. Figure 2.2 shows that in 2012, there was a large increase in Form ADV filings. Clear from the

³⁷See <https://www.law360.com/securities/articles/388275/sec-counts-1-500-fund-advisors-registered-under-dodd-frank>

figure is that, an increase in ADV-W filings drove the spike. To the extent some firms strategically relocate to avoid regulation, this would attenuate our estimates because these advisers are likely ones with higher misconduct records. However, strategically relocating is very costly given the changing client bases, since mid-sized advisers must register with any state in which they have business activity.

2.3.5 Summary Statistics

Table 2.1 displays the breakdown of observation counts for various subsets of the data. Across all person-years available in the data, we have 4.6 million observations, winnowed to 3.1 million for the 10-year period ending in 2016. This corresponds to 492,841 unique individuals and 30,579 unique firms in the IAPD database. The number of unique firms in the SEC data set through 2015 is 6,235, suggesting there are 24,000 firms that were never registered under the SEC. There are 1,791,522 person-year observations across SEC and state-registered advisers. Among the set of individuals working at once-SEC-registered firms available in the sample at the end of 2014, the number is 1.29 million person-year observations, corresponding to 382,000 observations in the two-period sample.³⁸

Table 2.2 presents summary statistics at the firm level. The number of treated firms and untreated firms is 2,316 and 3,910 respectively, which suggests that the change impacted more than 1/3rd of investment adviser firms. Percentile-size differences are imposed by the Act's cutoff of \$100 million. Along several dimensions, treated and untreated firms are similar. Both have similar levels of investment discretion and proprietary conflicts of interest.³⁹ However, compared to non-treated firms, treated firms are less likely to have custody of assets, less subject to independent audits, less likely to be private funds, less likely to recommend an external broker, and more likely to serve individuals and unsophisticated individuals. These compositional differences in

³⁸Conditional on getting a complaint, treated adviser representatives are not any likelier to leave the industry the next year. However, they are more likely to leave in the same year. Our recidivism findings suggest that this exit would bias our estimate of interest downwards. Finally, treated firms are not any likelier to shut down than untreated firms controlling for size. However, this does not take into account labor market mobility.

³⁹Potential proprietary conflicts of interests arise when an adviser's and clients' trading incentives may differ.

Table 2.1: Observation counts

The table below shows summary statistics of our data constructed by merging the universe of firms' Form ADV filings and the universe of individual adviser representative reports to create a survivorship bias free data set.

| Sample | N |
|---------------------------------------|-----------|
| # Unique representative CRDs in IAPD | 492,841 |
| # Unique firm CRDs in IAPD | 30,579 |
| # Unique CRDs in SEC Form ADV | 6,235 |
| # Firm-year observations in IAPD | 4,623,292 |
| # Obs (2009-2014) | 1,791,522 |
| # Obs (Annual Sample) | 1,290,043 |
| # Obs (Three-year-treatment window) | 382,665 |
| # Firms treated | 2,316 |
| # Individuals treated | 23,547 |
| # Individual-year treated | 129,171 |
| # Always State-registered firms | 17,821 |
| # Always State-registered individuals | 97,264 |

firm characteristics also justify the use of firm fixed effects. In some specifications, we go further to also include individual fixed effects. To further ease this concern, we also construct a matched sample comparison designed to maximize the similarity between individuals in each group to guard against any potential bias these observed differences may create.

Table 2.3 presents information on complaints before and after the Dodd-Frank Act. Panel A summarizes the probability of complaints and the dollar-value of alleged damages. The average probability of receiving a complaint in a given year is 1.25%. The complaint probability rises in bad times and declines in good times, evidenced by the drop in complaint rates from 1.63% immediately following the financial crisis, 2009-2012, to 0.85% during 2012-2015. Conditional on having alleged damages, the dollar amount of a complaint is significant with an average over \$200,000. Finally, the sample split shows that in the pre-period, both complaint rates and alleged damages were higher. We confirm that both groups have similar pre-trends visually (see Figure 2.5) and empirically in several ways.

Although not all complaints are arbitrated, they still reveal misconduct. Around a third of com-

Table 2.2: Firm Summary Statistics

The table provides summary statistics for investment adviser firms. All summary statistics are reported for the Form ADV filing for 2011. Firms check whether they have any custody over assets, independent audits, specific incentive structures, and their client compositions. They also report the fraction of assets that they have custody over.

| | Total | Untreated | Treated |
|--|-------|-----------|---------|
| N with non-missing AUM | 6,226 | 3,910 | 2,316 |
| Assets: | | | |
| AUM 10 th Percentile | 35 | 103 | 30 |
| AUM 25 th Percentile | 57 | 143 | 39 |
| AUM 50 th Percentile | 128 | 295 | 53 |
| AUM 75 th Percentile | 422 | 882 | 72 |
| AUM 90 th Percentile | 1,841 | 4,090 | 89 |
| Fraction of AUM with Custody (%) | 7 | 6 | 14 |
| Fraction of RIAs with Custody (%) | 18 | 24 | 8 |
| Fraction with Independent Audits (%) | 21 | 29 | 9 |
| Incentive Structure: | | | |
| Private Fund (%) | 26 | 32 | 15 |
| Other Business (%) | 16 | 15 | 18 |
| Other Business is Main Business (%) | 5 | 5 | 6 |
| Recommends a Broker (%) | 64 | 68 | 58 |
| Have Proprietary Conflicts of Interest (%) | 87 | 88 | 86 |
| Have Sales Conflicts of Interest (%) | 20 | 27 | 10 |
| Have Investment Discretion (%) | 93 | 94 | 90 |
| Client Composition: | | | |
| Individuals (%) | 69 | 66 | 75 |
| Unsophisticated Individuals (%) | 35 | 29 | 45 |
| Institutions (%) | 43 | 47 | 36 |

Table 2.3: Summary Statistics around Dodd-Frank

Panel A: The table below shows summary statistics for complaints in the pre- and post-treatment periods.

Complaints are filed against investment adviser representatives and employers, and may contain additional details such as alleged damages, associated products, complaint type, and outcomes. Observations are at the investment adviser representative-year level. There are less than 6,587 total firms in either the pre- or post-treatment period because firms leave the sample for reasons other than switching to file with the state.

| | Pre-Period: 2009 - 2012 | | | Post-Period: 2012 - 2015 | | |
|-----------------------------|-------------------------|-----------|---------|--------------------------|-----------|---------|
| | Total | Untreated | Treated | Total | Untreated | Treated |
| Complaints: | | | | | | |
| P(complaints) (%) | 1.63 | 1.65 | 1.05 | 0.85 | 0.85 | 0.80 |
| Complaints with Details (%) | 99 | 99 | 100 | 99 | 99 | 99 |
| Complaint Outcomes: | | | | | | |
| Withdrawn (%) | 52 | 53 | 42 | 50 | 50 | 38 |
| Settled (%) | 33 | 33 | 31 | 28 | 28 | 31 |
| Award or Arbitration (%) | 6 | 6 | 8 | 4 | 4 | 5 |
| Alleged Damages (\$) | 262,898 | 265,353 | 170,942 | 201,644 | 203,387 | 153,979 |
| Time to Resolution (days) | 155 | 155 | 154 | 117 | 117 | 128 |

Panel B: The below table breaks down complaints with details into those associated different products and complaint types. Complaint types are codified directly by the IAPD to classify complaints. Both the types of complaints and the products are not mutually exclusive. Customers may file complaints of multiple types, associated with multiple products.

| | Product | | | Type | |
|--------------------|---------|------------|----------------------|--------|------------|
| | Number | % of Total | | Number | % of Total |
| Annuity | 7,122 | 28.32 | Suitability | 9,955 | 39.58 |
| Variable Annuity | 6,498 | 25.84 | Misrepresentation | 8,620 | 34.27 |
| Mutual Fund | 5,371 | 21.36 | Fiduciary | 1,943 | 7.73 |
| Equity | 5,040 | 20.04 | Unauthorized Trading | 1,827 | 7.26 |
| Insurance | 2,108 | 8.38 | Fraud | 1,484 | 5.90 |
| Debt | 2,051 | 8.16 | Fees | 1,306 | 5.19 |
| Real Estate | 1,941 | 7.72 | Portfolio Allocation | 669 | 2.66 |
| OTC | 1,105 | 4.39 | Churning | 567 | 2.25 |
| Options | 591 | 2.35 | | | |
| Fixed Annuity | 578 | 2.30 | | | |
| Private Placements | 409 | 1.63 | | | |

plaints are settled and 5-8% of complaints result in arbitration or restitution. Panel B in Table 2.3 presents complaints by product and allegation type. In terms of product, complaints most fre-

quently involve annuities, mainly variable annuities, mutual funds, and equities. These are areas in which a financial adviser may be most useful. Mutual funds and annuities are products that require sophistication to navigate the large product space, while equities are highly informationally sensitive. In terms of allegation types, the most common complaints are for misrepresentation and lack of suitability. These standards are typically also standards against which brokers are held, but most advisers are brokers leading to an interpretation of the standards as violations of the more-stringent fiduciary standards. The explicit word “fiduciary” is present in about 7% of complaints, suggesting the client is alleging a violation of a broad fiduciary standard. The fourth largest category is “un-authorized”, also at around 7%, suggesting the adviser made a trade that the client did not authorize. There are relatively fewer complaints alleging “churning” and “fees,” which are broker related, because FINRA oversees brokers and FINRA was unaffected by the treatment.

2.4 Results

2.4.1 Complaint Incidence

Our baseline results in Table 2.4 compares the complaint rate for representatives at firms that switched from SEC to state-regulator oversight (Treated) to those at firms that remain registered with the SEC (Control). Regression (2) shows that the complaint rate for the treated group increased 50%, or 1.8 percentage-points, relative to the untreated group. This percentage is the probability of receiving a complaint in the next 3 years after 2012 - the treatment year. Regression (1) uses an annual panel and finds a 0.5-percentage-points increase in the complaint rate for the treated group, which is around 1/3 of the three-year effect in regression (2). The estimated treatment effect is stable across all specifications. Even in regression (5), in which we add firm and individual fixed effects, the increase in the complaint rate is 1.8 percentage-points. The results are similarly stable in regressions (6) to (9), which uses a log transformation of the number of complaints a representative receives in the three years pre- and post-2012.

Although our baseline results are significant and persistent, we run tests to confirm that the control group is valid. The concern is that larger investment advisers (Control) may not be comparable to smaller advisers (Treated), since larger advisers may engage in different business practices with different clienteles. Also, it may be possible that a positive estimate on $\text{Post}^{2012} \times \text{Treated}$ comes from a falling number of complaints for the control group rather than a rise in the complaints for the treated advisers. While firm and individual fixed effects remove unconditional differences between the treatment and control groups, the fixed effects do not completely rule out the concern that the treatment and control groups have different exposures to market conditions.

To mitigate these concerns, Table 2.5 considers a variety of robustness checks. First, in regression (3), we repeat the analysis comparing untreated-mid-size advisers headquartered in New York and Wyoming with treated-mid-size advisers. This test resolves concerns that compositional differences in advisers across mid-size and large-size advisers drive the results. For the second test, in regression (2), we exclude mid-size advisers headquartered in New York and Wyoming and only compare treated mid-size firms to untreated-large-size firms. This test ensures that compositional differences between treated-mid-size advisers and advisers in New York and Wyoming are not driving the results. In both the first and second tests, the response of advisers to treatment is remarkably similar - a 1.6-percentage-point increase for treated advisers relative to the control group. For the third test, we ensure that the results are not driven by outliers, removing California from the treatment group (the state with the largest number of advisers). In this case, we see an even larger treatment effect of a 4-percentage-point increase compared to the baseline of 1.8 percentage points. We also run a placebo test in regression (4), in which we define the treatment year as 2005. The placebo shows no statistically significant results and the point estimate is negative.

Another concern is a violation of the parallel trends assumption. To resolve this concern, we repeat the analysis using matched samples. For each representative treated in 2012, we find a matching representative. The first match in regression (5) requires that representatives have the same complaint history in the three-year window 2009 to 2011. Specifically, they are matched on their total cumulative complaint count, and whether they received a complaint in 2009, 2010 and

Table 2.4: Baseline Results

The table below shows the difference-in-difference estimates of the impact of switching from SEC to state registration on the propensity to receive a complaint. The sample is investment adviser representatives working at firms under SEC oversight (firms filing ADV to the SEC) in 2011. The first three columns present annual, person-year panels while the second three collapse three year period around 2012 as suggested by Bertrand et al. (2004). In the latter, all individuals must work the entire three year period to be considered. The dependent variable is the percent receiving a complaint. Robust standard errors clustered by state are presented.

| <i>Dependent variable:</i> | | $1_{\{Complaint_t\}} * 100$ | | | | |
|-------------------------------|----------------------|-----------------------------|----------------------------|----------------------|---------------------|----------------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Sample | Annual sample | | | Pre-post Sample | | |
| Constant | 1.266*** (0.168) | | | 3.693*** (0.479) | | |
| Post ²⁰¹² | -0.802*** (0.088) | | | -2.310*** (0.243) | | |
| Treated | -0.466*** (0.171) | | | -1.760*** (0.468) | | |
| Post ²⁰¹² ×Treated | 0.529*** (0.161) | 0.544*** (0.143) | 0.574*** (0.164) | 1.841*** (0.446) | 1.764*** (0.431) | 1.653*** (0.726) |
| Fixed Effects | | Firm + Year | Firm + Year +Individual | | Firm | Firm + Year +Individual |
| Observations | 1,299,819 | 1,299,819 | 1,299,819 | 382,665 | 382,665 | 382,665 |
| R ² | 0.003 | 0.011 | 0.252 | 0.007 | 0.024 | 0.628 |
| Adjusted R ² | 0.003 | 0.007 | 0.045 | 0.007 | 0.009 | 0.018 |

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

Table 2.5: Robustness: Alternative Comparison Groups

The table below shows alternative specifications. Column 1 removes California, which is home to the highest number of investment advisers. Column 2 excludes New York and Wyoming because advisers in these two states stayed with the SEC. Column 3 removes any adviser with over \$100 million of assets under management, forcing the control group to be other mid-size advisers in New York and Wyoming. Column 4 presents a placebo estimate defining the treatment year as 2005. Column 5 presents a matched sample exercise, where we find matches with replacement from the same state, with the same 3 year complaint count, and the same pre-trend (whether they received complaints in 2009, 2010, 2011). Finally, they are propensity score matched in the post period as working for a firm with the same probability of being treated. Column 6 is an alternative matched sample, relaxing the constraint that the individual comes from the same state. The unit of observation is at an individual level, in a 3 year window around the implementation of state registration. Robust standard errors clustered by state are presented in the OLS samples. Standard errors for the matched samples are clustered at the level of match pair following the recommendation of Abadie and Spiess (2016a).

| <i>Dependent variable:</i> | $1_{\{Complaint_t\}} * 100$ | | | | | |
|--------------------------------|-----------------------------|----------------------|----------------------|-------------------|----------------------|----------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Sample: | Exclude CA | Exclude NY/WY | ≤ 100 MM | Placebo | Match 1 | Match 2 |
| Post ²⁰¹² | -1.973*** (0.391) | -1.954*** (0.477) | -4.433*** (1.161) | 0.980 (1.212) | -1.648*** (0.162) | -1.978*** (0.175) |
| Post ²⁰¹² × Treated | 1.776** (0.764) | 1.634** (0.809) | 4.084*** (1.193) | -1.218 (1.342) | 1.807*** (0.200) | 1.729*** (0.213) |
| Fixed Effects | Firm + Individual | Firm + Individual | Firm + Individual | Firm + Individual | Firm + Individual | Firm + Individual |
| Observations | 362,982 | 277,158 | 20,065 | 225,434 | 107,240 | 108,810 |
| R ² | 0.631 | 0.648 | 0.727 | 0.659 | 0.651 | 0.655 |
| Adjusted R ² | 0.046 | 0.071 | 0.151 | -0.020 | 0.570 | 0.675 |

Note:

*p<0.1; **p<0.05; ***p<0.01

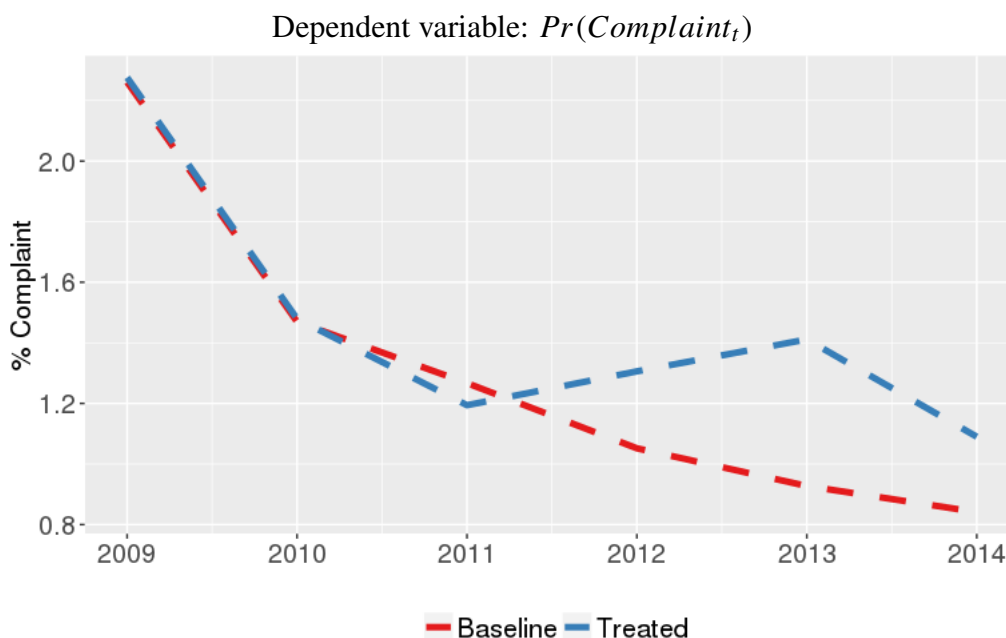


Figure 2.5: Parallel Trends

Figure shows trends in complaint rates for advisers registered with the SEC from 2009 to 2014 (baseline) with those advisers who switched from SEC to state regulation in 2012 (treated). The baseline group is formed based on a matched sample, forcing similar complaint rates in 2009 and 2010 and requiring advisers be in the same state. No complaint matching is done on 2011 data. The trend from 2010 to 2011 is parallel consistent with parallel trends but then diverges on treatment in 2012, when Dodd-Frank shifts oversight of mid-size (\$25-\$100M assets under management) advisers to state regulators.

2011. Representatives also have to operate in the same state and work at a firm with the same propensity of being treated, determined by a vector of firm characteristics. The matched sample regressions lead to a similar estimate of the treatment effect of 1.8-percentage-points. Regression (6) requires exact matches on year-by-year complaint counts pre-treatment and also confirms the estimate of the treatment effect. It relaxes the assumption of requiring firms to be from the same state. Figure 2.5 shows a parallel trend graph suggesting no violation in trends. For the graph, we matched advisers only on complaints in 2009 and 2010. Evidence in the graph, the matched samples have identical trends in 2011 and divergent trends on treatment in 2012.⁴⁰

We also use a regression discontinuity design and find confirming results in bands around \$100

⁴⁰In addition to the two main matching algorithms, we also considered other variants: using the same number of complaints total in history (instead of the pre-period), and forcing the same complaint *count* every year in the pre-period. All variants produce very similar results that are sometimes stronger. Other restrictions such as forcing zero complaints in the pre-period provide a directionally similar estimate. We also tried matching without replacement, achieving very similar magnitudes.

Table 2.6: Different AUM cutoffs

This table presents the main result using different AUM cutoffs. Firms with 2011 AUM above the stated amount are excluded, such that the control group is all untreated firms below the indicated AUM level. Cluster robust standard errors are clustered at the state level.

| <i>Dependent variable:</i> | $1_{\{Complaint_t\}} \times 100$ | | | | | | | |
|----------------------------|----------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| AUM Cutoff = | 100MM | 150MM | 200MM | 250MM | 300MM | 350MM | 400MM | 450MM |
| Past×Treated | 1.337*** (0.217) | 0.823*** (0.294) | 0.698** (0.272) | 0.640*** (0.236) | 0.491** (0.247) | 0.479** (0.218) | 0.469** (0.214) | 0.444** (0.210) |
| Fixed Effects | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual | Firm+Year+ Individual |
| Observations | 66,960 | 87,047 | 102,216 | 113,765 | 126,390 | 143,501 | 153,038 | 159,435 |
| R ² | 0.344 | 0.353 | 0.353 | 0.351 | 0.349 | 0.341 | 0.342 | 0.341 |
| Adjusted R ² | 0.053 | 0.069 | 0.075 | 0.074 | 0.076 | 0.071 | 0.075 | 0.076 |

*p<0.1; **p<0.05; ***p<0.01

million. However, the results are more noisy, trading off the closeness to the discontinuity threshold with the number of observations. Contributing to the lack to stability is that the density of firms immediately around the cutoff is low.⁴¹ Also, the discontinuity is not precise as advisers registered with the SEC do not need to de-register unless assets fall under \$90 million while newly registering advisers have to be above \$100 million to register with the SEC. However, our difference-in-difference test can be performed in a narrower window around the \$100 million cutoff. Table 2.6 presents the results. Sensibly, the point estimate is smaller moving further away from the cutoff. Interestingly, although the results become moderately less precise as we move away from the cutoff, double-clustering actually improves our results. However, we report single clustered results to remain consistent.

Taken together, we argue we have a robust finding: switching to state regulation resulted in a larger number of complaints for affected investment advisers. Additional robustness checks can be found in the Appendix.

2.4.2 Types of Complaints

The increase in complaints for treated representatives is mostly fiduciary-related, rather than broker-related. We expect this differential response in complaints because FINRA oversees brokers and FINRA was not affected by the treatment in 2012. Consistent with this logic, we find no significant increase in complaints related to churn (excessive trading), a broker-related misconduct. In contrast, fiduciary and unauthorized trading activity differentially increased after treatment.

The differential increase in complaints for the treated advisers is concentrated among options, equity, real estate, and private placements. The increase in complaints precipitates among equities and options, which are risky, complex and informationally sensitive assets that likely require the assistance of a financial adviser. We see null results for private placements, capturing activity with sophisticated clients (by the current legal definition, which is based on income and net worth), who at the margin are likelier to be able to govern their adviser. Annuities also yield a similarly null

⁴¹For example, firms are not required to report assets under management down to the dollar.

result, as annuities are buy-and-hold products and there is less ongoing involvement by an adviser or adviser representative advising a client after origination and sale.

2.4.3 Recidivism

We find that representatives most likely to take advantage of a weaker regulator do. The first test shows that regulators with greater histories of misconduct responded the most to treatment:

$$Complaint_i = \alpha + \beta_1 Past_i^{Complaint} + \beta_2 Treated + \beta_3 Past_i^{Complaint} \times Treated + \varepsilon_{it}.$$

Controlling for the unconditional recidivism rate, we test whether those with past complaints had a larger treatment effect after the re-jurisdiction. Table 2.8 reports the results. Past history can be defined using the number of complaints $\log(1 + \#Complaints)$, having a complaint $1_{\{Complaints\}}$, or the number of complaints relative to the expected number of complaints based on characteristics of the advisory firm, $\epsilon_t^{Complaints}$.⁴² Some specifications control for firm fixed effects, which controls for firm characteristics that predispose the representative to conflicts-of-interest and controls for selection into bad firms.

Table 2.8 regression (1) shows that representatives that had a complaint during 2009 to 2011 were 9.4% more likely to have a complaint in 2012 to 2015. The unconditional recidivism we observe is in line with Egan et al. (2016). For representatives at firms that switched to state regulation from SEC oversight, the probability of receiving a complaint during 2012 to 2015 increased 7.2% more, a 77% increase over the control group. In regression (2), we add firm fixed effects and find very similar results, suggesting that representatives that misbehaved more at a specific firm also responded more to the treatment than other representatives at the same firm. In regression (3), we limit the sample to only mid-size advisers to compare treated mid-size advisers to untreated mid-size advisers located in Wyoming and New York. We find directionally similar results and

⁴²The benchmark is a number of Form ADV characteristics indicating conflicts of interest and assets under management, plus state fixed effects. The residual comprises excess complaint variation not attributable to the firm or state.

the treatment response is even stronger.⁴³ Regressions (5) and (6) do tests using the number of complaints in the past and future and show that representatives that misbehaved relatively more also responded relatively more. Regressions (7) to (9) show a similar result but compare a representative's misbehavior to predicted misbehavior using firm and state characteristics.

2.5 Misconduct or Reporting?

We now examine whether the increase in complaints is due to more misconduct or reporting. The number of observed complaints is a function of actual misconduct and the probability of detection as well as the probability of a mistaken complaint and the amount of legitimate advisory activity:

$$Complaints = Pr(Detection) \times Misconduct + Pr(False\ Positive) \times Normal\ Activity.$$

Changes to the probability of detection of misconduct, would generate an increase in observed complaints. Also, if customers are more likely to file frivolous complaints, observed complaints would rise. Any combination of the two could generate the increase in complaints without any changes in misconduct. In order to distinguish between alternative driving forces behind the higher complaints, we rely on cross sectional tests.

2.5.1 Serious Complaints

If filing a complaint with a state regulator is perceived as less costly than filing a complaint with the SEC, then the increase in complaints for treated advisers may be due to more relatively less-severe complaints. However, there is nothing about the process of filing a complaint that suggests filing with a state regulator is less costly than filing with the SEC. Both venues allow online form submissions. To determine whether the increase in complaints is driven by more relatively-less-severe complaints, we analyze whether the dollar value of alleged damages and resolution outcomes are related to the treatment.

⁴³The results presented here winsorize past complaints. In some rare cases, past complaints total several dozens, reflecting the same complaint from multiple clients in a group-action type filing. These outliers attenuate our results rendering some results less precise (dropping results at the 5% level to the 10% level).

Table 2.7: Customer Complaint Decomposition

The table below decomposes the complaint types in the three-year-collapsed window around treatment in 2012. Dependent variables are scaled to be a probability multiplied by 100. All regressions include firm and year fixed effects. Robust standard errors clustered by state are presented.

| Panel A: Complaint Allegation Types | | | | | | | | | |
|-------------------------------------|----------------------|----------------------|--------------------|--------------------|----------------------|---------------------|----------------------|----------------------|----------------------|
| Allegation Type: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | Churning | Adviser | Portfolio | Fees | Fiduciary | Fraud | Misrepresentation | Suitability | Unauthorized |
| Post ²⁰¹² | -0.025*** (0.007) | -1.037*** (0.146) | -0.029* (0.017) | -0.108* (0.065) | -0.144*** (0.041) | -0.068** (0.031) | -0.792*** (0.216) | -0.855*** (0.255) | -0.101*** (0.019) |
| Post ²⁰¹² × Treated | 0.073 (0.076) | 0.600** (0.264) | 0.025 (0.028) | -0.047 (0.101) | 0.360 (0.257) | 0.363* (0.205) | 1.116*** (0.383) | 1.089** (0.420) | 0.111** (0.050) |
| Observations | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 |
| R ² | 0.012 | 0.020 | 0.035 | 0.007 | 0.043 | 0.040 | 0.026 | 0.024 | 0.006 |
| Adjusted R ² | -0.002 | 0.007 | 0.022 | -0.007 | 0.034 | 0.027 | 0.012 | 0.010 | -0.008 |

| Panel B: Complaint Product Type | | | | | | | | | |
|---------------------------------|-------------------|----------------------|----------------------|----------------------|--------------------|--------------------|------------------|--|--|
| Product Type: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | | |
| | OTC | Options | Insurance | Equity | Real Estate | Private Pl | Managed Act | | |
| Post ²⁰¹² | -0.013 (0.018) | -0.018*** (0.006) | -0.169*** (0.048) | -0.333*** (0.051) | 0.057** (0.026) | -0.048* (0.026) | 0.011 (0.010) | | |
| Post ²⁰¹² × Treated | 0.139 (0.119) | 0.093** (0.033) | -0.394* (0.225) | 0.409** (0.192) | 0.409** (0.189) | 0.063 (0.042) | 0.007 (0.018) | | |
| Observations | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | 381,006 | | |
| R ² | 0.018 | 0.006 | 0.015 | 0.014 | 0.034 | 0.029 | 0.011 | | |
| Adjusted R ² | 0.004 | -0.007 | 0.002 | 0.001 | 0.021 | 0.016 | -0.003 | | |

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

Table 2.8: The Effect of Regulatory Jurisdiction on Recidivism

This table presents a cross-sectional test of complaints received in the 2012-2014 period. The variable of interest is $Past \times Treated$, which is the interaction term between a variable describing the individual's past complaint history. Also of interest is the unconditional recidivism term, $Past$, which measures the amount of recidivism for complaints over the sample period. The measures of past activity include whether the individual has received a complaint before $1_{\{Past\ Complaints \geq 0\}}$, the log number of complaints $\log(1 + \#Past\ Complaints)$, and the residual number of complaints ϵ^{Past} . The latter is the residual of a regression based on a benchmark based on the characteristics of firms in which the individual worked in the past. Robust standard errors clustered by state are shown in parentheses.

| <i>Dependent variable:</i> | | $1_{\{Complaint_t\}} \times 100$ | | | | | | | | |
|----------------------------|---------------------|-----------------------------------|----------------------|--------------------|---------------------------------|---------------------|----------------------|----------------------|----------------------|--|
| Past Measure = | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | |
| | | $1_{\{Past\ Complaints \geq 0\}}$ | | | $\log(1 + \# Past\ Complaints)$ | | | ϵ^{Past} | | |
| Constant | 2.409*** (0.037) | | 1.402*** (0.232) | | | | | 1.380*** (0.053) | | |
| Past | 9.346*** (0.367) | 8.658*** (0.369) | 4.800** (2.137) | 4.752** (2.370) | 12.139*** (0.533) | 6.658* (3.413) | 30.206*** (1.095) | 27.841*** (1.181) | 15.885** (7.567) | |
| Treated | -0.023 (0.187) | | 1.032*** (0.298) | | | | -0.269 (0.253) | | | |
| Past×Treated | 7.145*** (2.514) | 4.820* (2.513) | 11.390*** (3.277) | 8.465** (3.637) | 7.114** (3.626) | 12.139** (5.243) | 19.907*** (6.944) | 15.581* (7.963) | 27.003** (11.622) | |
| Fixed Effects | | Firm | | Firm | Firm | Firm | | Firm | Firm | |
| Sample | Full | Full | ≤100 MM | ≤100 MM | Full | ≤100 MM | Full | Full | ≤100 MM | |
| Observations | 183,423 | 183,423 | 9,723 | 9,723 | 183,423 | 9,723 | 183,423 | 183,423 | 9,723 | |
| R ² | 0.014 | 0.038 | 0.028 | 0.185 | 0.040 | 0.186 | 0.016 | 0.038 | 0.185 | |
| Adjusted R ² | 0.014 | 0.009 | 0.027 | -0.034 | 0.010 | -0.033 | 0.016 | 0.009 | -0.034 | |

* p<0.1; ** p<0.05; *** p<0.01

Table 2.9 shows that the dollar value of alleged damages increases by between \$103,000 to \$105,000 for the treatment group. When including time trends in overall complaints and firm-specific complaints, complaints had \$105,678 more alleged damages. If complaints become more frivolous, the average alleged dollar value of damages for complaints should be lower for the treatment group when Dodd-Frank comes into effect. This is not supported by the data. If anything, damages increase. It could be, however, that these complaints with higher-alleged damages are also less-well founded and thus more likely to be dismissed. Panel B of Table 2.10 suggests otherwise. Looking at whether complaints later get withdrawn or closed with no action, regressions 1 and 2 show a negative, but statistically insignificant, coefficient on the interaction term. This suggests that the withdrawal rate did not change. If anything, it decreased. Regression 3 and 4 show that follow-up legal outcomes did not change. When complaints were not withdrawn, or closed with no action, or settled with no admission of guilt, they were not more likely to be denied in arbitration.

The nature of complaints do not seem to be materially different than before. Regressions 1 and 2 in Panel A of Table 2.10 show that complaints did not take longer to resolve. Regressions 3 and 4 focuses on differences in arbitration settlement amounts as a proportion of dollars of alleged damages (e.g., a \$25,000 arbitration award based on \$100,000 alleged damages is a settlement ratio of 25%). If claims were less well founded, then settlement amounts should also decline relative to alleged damages. Instead, regressions (4) to (6) show that the settlement amounts do not decrease but if anything seem to increase for complaints against treated advisers. However, if complaints were less serious, the average time to resolution should decrease, and regressions (1) to (3) are consistent with the time to resolution decreasing.

Altogether, there is no significant evidence that the complaints filed against treated advisers became less severe. Complaints had higher alleged damages, but were not more or less likely to be withdrawn, dismissed, or denied in arbitration. Therefore, the higher incidence of complaints are not driven only by more frivolous complaints.

Table 2.9: Alleged Damages Analysis

The table below shows results for alleged damages for client complaints. Alleged damages are in dollars. Observations are at the investment adviser representative by year level. Robust standard errors clustered by firm headquarter state are shown in parentheses.

| <i>Dependent Variable:</i> | Alleged Damages | |
|-------------------------------|-------------------------------|-----------------------------|
| | (1) | (2) |
| Post ²⁰¹² | -117,933.90*** (33,218.34) | |
| Post ²⁰¹² ×Treated | 103,542.90*** (44,645.19) | 105,678.40** (53,032.42) |
| Fixed Effects | Firm | Firm + Year |
| Observations | 33,992 | 33,992 |
| R ² | 0.009 | 0.010 |

*p<0.1; **p<0.05; ***p<0.01

Table 2.10: Complaint Noise-to-Signal Ratio

Panel A: The table below shows results for time to resolution and the monetary compensation relative to alleged damages for customer complaints. Observations are at the investment adviser representative by year level. Robust standard errors clustered by IAR state are shown in parentheses.

| <i>Dependent Variable:</i> | (1) | (2) | (3) | (4) |
|-------------------------------|-----------------------|---------------------|--|------------------|
| | Time to Resolution | | $\frac{\$ \text{ Award Amount}}{\$ \text{ Alleged Damages}}$ | |
| Post ²⁰¹² | -40.776** (18.701) | | -6.085 (4.776) | |
| Post ²⁰¹² ×Treated | -7.625 (38.104) | -24.509 (41.021) | 5.856 (4.776) | 4.259 (3.687) |
| ln (\$Alleged Damages) | 2.718** (0.340) | 1.565*** (0.808) | | |
| Fixed Effect | Firm | Firm + Year | Firm | Firm + Year |
| Observations | 33,098 | 33,098 | 7,355 | 7,355 |
| R ² | 0.100 | 0.133 | 0.010 | 0.022 |

*p<0.1; **p<0.05; ***p<0.01

Panel B: The table below shows the complaint outcome. $I_{\{\text{Withdrawn/No Action}\}}$ is an indicator of whether the complaint was later withdrawn or no further follow-up action occurred. $I_{\{\text{Denied}\}}$ is an indicator of whether the complaint was denied, conditional on it not being “Withdrawn” or “Closed/No Action”. Robust standard errors clustered by IAR state are shown in parentheses.

| | (1) | (2) | (3) | (4) |
|-------------------------------|--------------------------------------|-------------|-------------------------|-------------|
| <i>Dependent Variable:</i> | $I_{\{\text{Withdrawn/No Action}\}}$ | | $I_{\{\text{Denied}\}}$ | |
| Post ²⁰¹² | -0.038** | | 0.070** | |
| | (0.018) | | (0.008) | |
| Post ²⁰¹² ×Treated | -0.037 | -0.056 | -0.047 | -0.020 |
| | (0.038) | (0.037) | (0.049) | (0.049) |
| ln (\$Alleged Damages) | -0.004*** | -0.005*** | -0.027*** | -0.024*** |
| | (0.001) | (0.001) | (0.003) | (0.003) |
| Fixed Effect | Firm | Firm + Year | Firm | Firm + Year |
| Observations | 28,128 | 28,128 | 16,594 | 16,594 |
| R ² | 0.084 | 0.102 | 0.202 | 0.022 |

*p<0.1; **p<0.05; ***p<0.01

2.5.2 Under-staffed State Regulators

States that are more financially constrained are less able to investigate and deter misconduct. In response, representatives may take advantage of a weaker regulator by misbehaving more. Figure 2.6 shows that although state regulators’ workload increased significantly, on average they did not increase their requested budgets. To better identify the impact of regulator budget on observed misconduct, we study heterogenous treatment effects across states with different funding.⁴⁴

Our main measure of state regulator resources is staff-per-adviser. To measure staff-per-adviser, we use a report compiled in 1999 by the American Association of Retired Persons on the regulatory differences for investment advisers in every state. We argue that the staff-per-adviser in 1999 devoted to adviser regulation is likely correlated with the current level of regulatory oversight. Because it is predetermined, it is not contaminated by reverse causality.⁴⁵ In some states, the staff-per-adviser number is not available, which is attributable to the fact that the organization of the

⁴⁴ Another (albeit contrived) interpretation might be that in fiscally constrained states, pension promises are less credible, demand for investment advisory services are greater, resulting in a greater strategic response.

⁴⁵ As part of this project, we tried surveying regulators today for the same numbers, but as obtaining pre-2012 numbers was extremely difficult. Even present day numbers are difficult.

Table 2.11: Staffing of the Investment Adviser Regulatory Office

This table presents the difference-in-difference estimates with triple interactions on the staff-per-regulated-firm devoted to adviser regulation in the year 1999. Staff-per-Firm is the number of oversight employees at the state regulator divided by the number of registered investment adviser firms. We drop the state with missing staff numbers. Robust standard errors clustered by state are shown in parentheses, except in the matched sample where it is clustered at the level of a matched pair as recommended by Abadie and Spiess (2016a).

| <i>Dependent variable:</i> | $1_{\{Complaint_t\}} \times 100$ | |
|---|----------------------------------|----------------------|
| | (1) | (2) |
| Post ²⁰¹² | -2.116*** (0.176) | -2.423*** (0.290) |
| Post ²⁰¹² × Staff-per-Firm | -0.052 (0.184) | -0.240 (0.165) |
| Post ²⁰¹² × Treated | 1.222*** (0.218) | 1.396*** (0.428) |
| Treated × Staff-per-Firm | -53.191 (33.809) | 0.664*** (0.137) |
| Post ²⁰¹² × Treated × Staff-per-Firm | -1.249*** (0.291) | -0.840*** (0.317) |
| Fixed Effects | Firm + Year | Firm + Year |
| Sample | ≤100MM | Annual |
| Observations | 88,761 | 337,472 |
| R ² | 0.139 | 0.024 |
| Adjusted R ² | 0.101 | 0.009 |

*p<0.1; **p<0.05; ***p<0.01

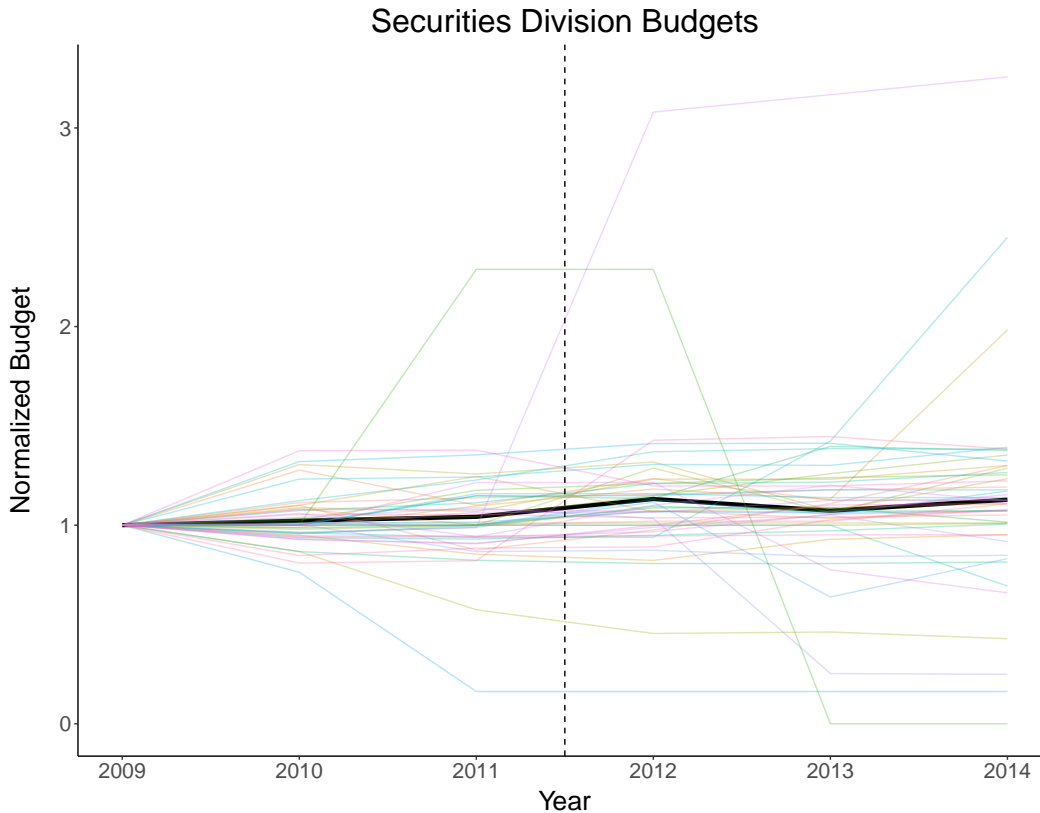


Figure 2.6: State Regulator Budgets

The figure below shows the proposed state regulator budgets, where the requested budget for calendar year 2009 is normalized to be 1 due to different state departmental organizations. The bolded line is the average across all states. Most states request budgets on a biannual basis.

regulatory body is *not* devoted to adviser regulation only, but financial services regulation overall.

Table 2.11 shows that states with more staff-per-adviser saw less of a treatment response, consistent with more staff at securities regulators deterring more misbehavior. The first four specifications impute a value of zero staff for those states with missing values. The last two columns drop states that may have no staff devoted specifically to adviser regulation, which does not qualitatively change the results. The equivalent point estimates (1) versus (5) across the two specifications report slightly lower treatment magnitudes when considering only states with designated staff. The relative magnitudes suggest that the treatment effect is larger if the state had no staff overseeing advisers more than a decade before treatment, which is sensible.

2.5.3 Dilution of Regulatory Resources Impacting Incumbent Advisers

Prior to Dodd-Frank, state regulators oversaw small-adviser firms, managing less than \$25 million in assets. On one hand, if state regulators became less effective due to the added requirement to oversee mid-size advisers (\$25-\$100 million of AUM), then misbehavior by small-size advisers should increase. On the other hand, if complaints are driven by the monitoring ability of the regulator and Dodd-Frank increases the workload for the state-securities regulator, then complaint rates against small-size advisers should decrease. Table 2.12 regression (1) shows that always-state-registered firms saw a statistically significant increase in complaint rates of 0.51 percentage-points. The effect is robust to specifications using individual and firm fixed effects and matched samples, where the match was again done on the pre-trend.⁴⁶ This increase in complaint rates by always-state-registered advisers is consistent with representatives at small-size advisers responding strategically to a weaker regulatory environment by increasing misbehavior.

2.5.4 Distance from Regulators Impacting Oversight

Another way to show that misbehavior increased for advisers with the weaker regulatory oversight is to show that advisers that are harder or costlier to monitor responded more to treatment. If complaints increased because state regulators are better monitors, then increases should be greater for treated advisers located closer to the relevant securities regulator. Advisers located closer to a regulator are easier to travel to and learn about.

To measure distance, we use the longitude and latitude of the zip code of the branch where the representative worked and the location of the representative's branch, firm headquarters, nearest SEC office, and nearest FINRA office. Distances are measured using zip code coordinates from the 2013 Census, although we try other geocoding measures based on Google Maps and the Bing Maps API.⁴⁷

⁴⁶Since we do not observe the ADV data for the never-SEC-registered firms, we cannot do propensity score matching on which firm characteristics affect the probability of being treated. Instead we do nearest neighbor matching on historical complaint count and number of years as an adviser representative.

⁴⁷The distance measurements are quite similar and results were directionally and quantitatively similar.

Table 2.12: Treatment Effect on Existing State-Registered Firms

The table below shows the treatment effect on firms with who were always state registered (and never registered with the SEC) which had $AUM < \$25M$. The specifications are either a matched sample on individual characteristics, an annual panel, or a pre-post collapsed sample a la Bertrand et al. (2004). Dependent variables are in scaled to be a probability multiplied by 100. Robust standard errors clustered by state are shown in parentheses, except in the matched sample where it is clustered at the level of a matched pair as recommended by Abadie and Spiess (2016a).

| <i>Dependent Variable:</i> | $1_{\{Complaint\}} \times 100$ | | | |
|--|--------------------------------|--------------------|-------------------|-----------------------|
| | (1) | (2) | (3) | (4) |
| Post ²⁰¹² | | | | -0.9688** (0.1292) |
| Post ²⁰¹² × Always State-Registered | 0.512*** (0.093) | 0.837*** (0.25) | 0.694* (0.393) | 0.5766** (0.213) |
| Specification | Annual | Pre-Post | Pre-Post | Match |
| Observations | 1,732,674 | 662,426 | 662,426 | 403,924 |
| Fixed Effects | Firm + Year | Firm | Firm + Individual | Firm |
| R ² | 0.018 | 0.036 | 0.621 | 0.091 |
| Adjusted R ² | 0.006 | 0.005 | -0.021 | 0.061 |

* p<0.1; ** p<0.05; *** p<0.01

Table 2.13: Distance to Regulator

This table presents difference-in-difference estimates using the annual panel data. Local offices addresses are as of 2015 from the websites of NASAA, FINRA and the SEC. Distances in miles are calculated using coordinates of the zip code of the firm or regulator's address. Robust standard errors clustered by state are shown in parentheses.

| <i>Dependent variable:</i> | | $1_{\{Complaint_t\}}$ | | | | | |
|--|---------------------|-----------------------|---|-------------------------------------|--|--|--|
| | (1) | (2) | (3) | (4) | (5) | (6) | |
| Post ²⁰¹² ×Treated | 0.580*** (0.139) | 0.393*** (0.148) | 0.015 (0.231) | -0.025 (0.244) | 0.046 (0.235) | 0.260 (0.181) | |
| Post ²⁰¹² ×log(<i>Dist_{State}</i>) | -0.005 (0.086) | 0.043 (0.069) | -0.245 (0.296) | -0.446 (0.359) | -0.253 (0.276) | 0.152 (0.106) | |
| Post ²⁰¹² ×log(<i>Dist_{FINRA}</i>) | | | 0.105 (0.173) | 0.049 (0.198) | 0.107 (0.146) | 0.105 (0.163) | |
| Post ²⁰¹² ×log(<i>Dist_{SEC}</i>) | | | -0.065 (0.247) | 0.016 (0.280) | -0.040 (0.222) | -0.038 (0.215) | |
| Post ²⁰¹² ×Treated×log(<i>Dist_{State}</i>) | 0.276* (0.143) | 0.254** (0.109) | 0.382** (0.169) | 0.411** (0.172) | 0.377** (0.160) | 0.338** (0.170) | |
| Post ²⁰¹² ×Treated×log(<i>Dist_{FINRA}</i>) | | | 0.307 (0.206) | 0.334 (0.219) | 0.364* (0.216) | 0.149 (0.192) | |
| Post ²⁰¹² ×Treated×log(<i>Dist_{SEC}</i>) | | | -0.154 (0.372) | -0.284 (0.432) | -0.084 (0.393) | -0.233 (0.262) | |
| Fixed Effects | Firm Year | Firm State×Year | Firm <i>HQ Zip</i> ×Post ²⁰¹² | Firm Branch×Post ²⁰¹² | Firm <i>HQ Zip</i> ×Post ²⁰¹² Branch×Post ²⁰¹² | Firm <i>HQ Zip</i> ×Post ²⁰¹² Branch×Post ²⁰¹² Individual | |
| Observations | 1,164,033 | 1,164,033 | 1,164,033 | 1,116,530 | 1,164,033 | 1,164,033 | |
| R ² | 0.010 | 0.011 | 0.011 | 0.024 | 0.034 | 0.272 | |
| Adjusted R ² | 0.006 | 0.007 | 0.005 | 0.009 | 0.011 | 0.044 | |

*p<0.1; **p<0.05; ***p<0.01

Table 2.14: Client Composition

This table presents the difference-in-difference estimates sorted by client composition of the firms. The Private Fund indicator takes a value of 1 if the investment adviser is a private equity, venture capital, or hedge fund fund, which can only take money from high net-worth clients, institutions and governments. Regressions 2 through 7 use the majority type of clients that the investment adviser has based on answers to Form ADV Section 5D1 (a)-(m). High net worth individuals are classified as sophisticated individuals. Form ADV include . All specifications include firm and year fixed effects, subsuming the Post²⁰¹², Treated, and Client Composition variables. Robust standard errors clustered by firm headquarter state are shown in parentheses.

| <i>Dependent variable:</i> | | $1_{\{Complaint_t\}} \times 100$ | | | | | |
|---|--------------------|----------------------------------|---------------------|--------------------------------|------------------------------|------------------|---------------------|
| Fund/Client Type: | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| | Private Fund | Government | Institutions | Unsophisticated Individuals | Sophisticated Individuals | Individuals | Not Individuals |
| Post ²⁰¹² ×Treated | 0.484** (0.213) | 0.544*** (0.143) | 0.554*** (0.158) | 0.338 (0.313) | 0.538*** (0.160) | 0.297 (0.479) | 0.554*** (0.158) |
| Post ²⁰¹² ×Client Type | -0.130 (0.183) | 1.298*** (0.443) | 0.346*** (0.119) | -0.106 (0.214) | 0.283** (0.121) | 0.057 (0.307) | 0.351*** (0.119) |
| Post ²⁰¹² ×Treated×Client Type | -0.003 (0.405) | -0.995** (0.475) | -0.261 (0.229) | 0.294 (0.365) | -0.135 (0.250) | 0.321 (0.507) | -0.266 (0.230) |
| Fixed Effects | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year |
| Observations | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 |
| R ² | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| Adjusted R ² | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |

*p<0.1; **p<0.05; ***p<0.01

Table 2.13 shows that complaint rates increased more for treated firms located further from their appropriate state regulator. Regression (1) shows a positive and significant coefficient on the interaction of treatment and distance using firm and year fixed effects. An adviser one-standard-deviation farther away from a regulator (181 miles farther) has a 50% higher chance of receiving a complaint. The point estimates are robust to adding state-by-year fixed effects, zip-by-year fixed effects, and firm-individual-branch-zip fixed effects. These increasingly stringent specifications rule out other drivers of the relation like local economic conditions.

Each specification also includes interactions on the relation between treatment and distance from the closest office of FINRA and the SEC. In all specifications, being far from the relevant state regulator is primarily responsible for the observed relation between increased complaint rates and treatment. The distance from a FINRA branch also receives a large point estimate. Clients may complain to FINRA offices when misconduct increases. The distance from the nearest SEC offices is not significantly related to treatment and the coefficient is negative, if anything.⁴⁸

2.5.5 Client Composition and Treatment Response

We find that complaints increased the most for firms whose clients are less sophisticated. Following Egan et al. (2016), we look at client composition at the branch level. Individual adviser representatives are not required to report their client composition. However, we assume that client composition is likely correlated to the demographics of the county in which the adviser serves. We obtain the branch location (city level) in which the adviser works and assign the adviser to a county based on the city name.⁴⁹ After doing so, we then obtain county-level characteristics from various

⁴⁸Because investment advisers representatives can move across different firms throughout the sample, we emphasize annual specifications for this result. Although we report results using a full annual panel, using the $\leq \$100M$ AUM adviser panel yields even stronger results in terms of point estimates. Two-period results point in similar directions, but are weaker unless taking into account the fact that adviser representatives move across firms. Finally, a purely cross-sectional test, wherein we control for treatment, past complaints, and $\log(\text{distance})$, we also find that there is a significant relation between $\text{Treated} \times \log(\text{Distance})$.

⁴⁹After cleaning branch-city names for misspellings, we assign the city name to all relevant zip codes. Where a branch-name could correspond to multiple counties, we conservatively assign the adviser representative to the largest county. The vast majority of adviser representatives report a branch. However, some observations are lost due to a lack of data, or a branch city location that can not be disambiguated. In some cases, the adviser representative reports a branch location that is a state or an incomplete city name that does not correspond to a identifiable county.

Table 2.15: Client Composition

This table presents the difference-in-difference estimates sorted by client composition of the firms. The Private Fund indicator takes a value of 1 if the investment adviser is a private equity, venture capital, or hedge fund fund, which can only take money from high net-worth clients, institutions and governments. Regressions 2 through 7 use the majority type of clients that the investment adviser has based on answers to Form ADV Section 5D1 (a)-(m). High net worth individuals are classified as sophisticated individuals. Form ADV include . All specifications include firm and year fixed effects, subsuming the Post²⁰¹², Treated, and Client Composition variables. Robust standard errors clustered by firm headquarter state are shown in parentheses.

| <i>Dependent variable:</i> | | $1_{\{Complaint_t\}} \times 100$ | | | | | |
|---|--------------------|----------------------------------|---------------------|--------------------------------|------------------------------|------------------|---------------------|
| Fund/Client Type: | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| | Private Fund | Government | Institutions | Unsophisticated Individuals | Sophisticated Individuals | Individuals | Not Individuals |
| Post ²⁰¹² ×Treated | 0.484** (0.213) | 0.544*** (0.143) | 0.554*** (0.158) | 0.338 (0.313) | 0.538*** (0.160) | 0.297 (0.479) | 0.554*** (0.158) |
| Post ²⁰¹² ×Client Type | -0.130 (0.183) | 1.298*** (0.443) | 0.346*** (0.119) | -0.106 (0.214) | 0.283** (0.121) | 0.057 (0.307) | 0.351*** (0.119) |
| Post ²⁰¹² ×Treated×Client Type | -0.003 (0.405) | -0.995** (0.475) | -0.261 (0.229) | 0.294 (0.365) | -0.135 (0.250) | 0.321 (0.507) | -0.266 (0.230) |
| Fixed Effects | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year | Firm + Year |
| Observations | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 | 1,299,819 |
| R ² | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| Adjusted R ² | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |

*p<0.1; **p<0.05; ***p<0.01

Table 2.16: Client Composition at the Branch Level

This table presents the difference-in-difference estimates interacting treatment with client demographics in the county in which the adviser representative works. The variable of interest is the triple interaction between treatment in the post period and a relevant demographic characteristic from the 2012 1-year ACS. % College is defined as the fraction of the county's adult population from age 24-54 with at least a Bachelor's education, while % Age > 60 is the fraction of the population aged 60 or above. Robust standard errors clustered by state are shown in parentheses.

| <i>Dependent variable:</i> | $1_{\{Complaint_t\}} \times 100$ | | | |
|--|----------------------------------|-----------------------|-----------------------|-----------------------|
| | (1) | (2) | (3) | (4) |
| County Type: | % College | % College | % College | % Age > 60 |
| Post ²⁰¹² × Treated | 0.578** (0.166) | 0.567*** (0.166) | 0.560*** (0.185) | 0.563*** (0.182) |
| Post ²⁰¹² × County Type | 0.194*** (0.183) | 0.212*** (0.443) | 0.0721 (0.117) | -0.057** (0.028) |
| Post ²⁰¹² × Treated × County Type | -0.187** (0.093) | -0.194** (0.093) | -0.232** (0.0103) | 0.345*** (0.092) |
| Fixed Effects | Firm + Year | Firm + Year Branch | Firm + Year Branch | Firm + Year Branch |
| Observations | 1,232,838 | 1,232,838 | 1,179,899 | 1,180,101 |
| R ² | 0.011 | 0.022 | 0.020 | 0.020 |
| Adjusted R ² | 0.006 | 0.010 | 0.007 | 0.010 |

*p<0.1; **p<0.05; ***p<0.01

government sources.

Table 2.16 reports our results. We use two versions of the American Community Survey. Although there is a minor amount of look-ahead bias, we first use the 2011-2015 American Community Survey as reported by the United States Department of Agriculture. Second, we use the American Community Survey 1-year survey from 2012, which has slightly worse coverage at the county-level for our sample. Columns 1 and 2 interact the treatment effect with county-level college degree attainment. It shows that a standard deviation increase in the county-level percentage of adults without college degrees increases treatment by 18.7 basis points, or about 1/3rd the unconditional mean. Column 2 shows adds a branch fixed effect. Controlling for county-level unconditional variation *increases* the point estimate to 19.4 basis points. Column 3 reports results based on the 2012 survey. It suggests a similar result: a 1 standard deviation increase in the fraction of population with bachelor degrees decreases treatment by 23.2 basis points. The fourth column shows that the treatment effect is greatest in counties with a larger fraction of elderly. However, slightly inconsistent with Egan et al. (2016), we do not observe a reliable relation between treatment and ethnic demographics.

We also assess client sophistication using reported client compositions by advisers on their Form ADV filings. We measure the fraction of clients who are less sophisticated with the proportion of investors who are unaccredited (net worth less than \$1 million or less than \$300,000 in income). Table 2.15 shows that complaint rates increased more when clients were unsophisticated and less when clients were sophisticated investors. Categories of more-sophisticated investors include accredited investors, government institutions, and institutional investors. These more-sophisticated investors are likely more capable at monitoring their investment advisers. However, these results are not statistically significant.

2.6 Conclusion

We show evidence that regulation deters misconduct by documenting higher complaints for firms facing a weaker regulatory regime. Specifically, we find that the re-jurisdiction of firms from SEC

to state regulators due to the Dodd-Frank Act increased misconduct by 50% on average. Firms in states with less regulatory resources and those that are further away from the state regulator saw the highest increases in complaints. Investment adviser representatives with poor prior track records do worse conditional on treatment. Overall, we estimate the cost in terms of higher alleged damages to be between \$10 and \$14 million. Counting increases in complaint incidence, conservatively, the attributable alleged damages hover around \$70 million.

Although client sophistication is able to alleviate this effect to some degree, our suggest regulation still has a role for governance of investment advisers in the United States. Our findings suggest that state securities regulators - while well-meaning - may potentially be less of a deterrent than federal regulators on average. If given only the choice of the state or national regulator, the results suggest that a Pareto gain would be made by allocating the same resources states have for regulation to the SEC.⁵⁰

More broadly, our results provide evidence the regulator matters above and beyond the writ of regulation itself, buttressing the thrust of Agarwal et al. (2014). In addition, we lay the foundation for future studies to study the governance of investment advisers by documenting the importance of regulator involvement and deterrence in understanding why conflicts between advisers and clients arise. Future research can study how different regulators implement the same law change, how local securities laws substantively differ between states and how that influences product market outcomes, and other institutional details that exacerbate agency issues when a firm advises clients as a fiduciary while also serving them as a broker.

⁵⁰It is possible that some states are in fact better than the SEC or no worse; however, we only identify the average effect.

Chapter 3

Speech is Silver but Silence Is Golden: Information Suppression and the Promotion of Innovation

3.1 Introduction

It is well-understood that information frictions inhibit financial markets from efficiently allocating capital towards innovative firms and projects (Hall and Lerner, 2010). Disclosure to investors helps to alleviate these frictions and is considered a fundamental pillar of modern capital markets. However, especially in the case of innovative activities, disclosure does not unambiguously serve the common good. In his seminal work on the welfare economics of knowledge production, Kenneth Arrow (Arrow, 1962) put forth the paradox that in order to determine the economic value of an idea, the inventor is forced to disclose information about the idea, and any such action inherently reduces its value. Indeed, Arrow's disclosure paradox has implicitly been at the heart of recent debates on the design of emerging forms of financing for startups, particularly crowdfunding, in which the requirement to disclose information to investors poses a risk for entrepreneurs in terms of eroding their potential competitive advantage.⁵¹ It has also influenced calls by policymakers and scholars to encourage more transparent record-keeping on the activities of innovative firms.⁵² We provide new empirical evidence on Arrow's disclosure paradox in the context of intellectual property (IP) licensing, a setting in which the investors' need for disclosure comes into conflict with the innovating firms' need for secrecy.

IP licensing is an economically important activity to study, unto itself. IP is perhaps the pri-

⁵¹In a recent interview, Josh Lerner notes "A lot of my doubts have to do with the inherent contradictions between the entrepreneurial process and disclosure requirements [...] but the very process of disclosing things is likely to destroy a lot of the competitive advantage that the entrepreneurs might have". For text of the full interview, see https://www.richmondfed.org/publications/research/econ_focus/2016/q2/interview, retrieved 3/11/2017.

⁵²In a recent open letter to the SEC Chien et al. (2016), several legal and innovation scholars including the USPTO chief economist request the SEC to make data on IP disclosures, which we study in this chapter, more available for public scrutiny. In 2009, the SEC started requiring firms to electronically file Form D, causing some practitioners to comment that the disclosure requirement is making it harder for young, innovative firms to raise capital. For more details, see <http://redeye.firstround.com/2008/06/the-death-of-st.html>, retrieved 3/11/2017.

mary asset for innovative firms. For such firms to generate economic value, they must be able to protect and commercialize their IP assets. Firms commercialize IP either by developing their own products and services or by trading these assets in the market for ideas (Gans and Stern, 2003). IP licensing agreements represent the main form by which firms commercialize their IP outside the firm (Shapiro, 1985), with total IP licensing royalty revenues of US corporations amounting to \$195 billion in 2013.⁵³ Licensing allows for an efficient division of labor between upstream firms or non-commercial entities (governments or universities) which are better suited for conducting research, and downstream firms with product market expertise, avoiding duplicative investments in marketing, manufacturing, and R&D programs.

This vibrant marketplace for ideas offers an ideal setting for examining Arrow's disclosure paradox. Publicly traded firms are required to disclose material IP licensing agreements.⁵⁴ By the definition of "materiality", these contracts are of significant economic value to investors. However, because they contain information on the value of ideas, their disclosure could result in competitive harm to the filing firms by revealing sensitive information to both industry incumbents and potential entrants. Wary of this issue, the US Securities and Exchange Commission (SEC) allows firms to file for "Confidential Treatment Requests" (CTR), granting firms the right to redact information from their mandatory filings that would constitute "competitive harm" to the filer. This is equivalent to non-disclosure of the underlying technology and the potential rents from licensing it.

When deciding whether or not to redact information, innovative firms balance the potential benefits of disclosure to investors with the potential costs of revealing information to rivals. However, firms do not face uniform costs and benefits of redaction. Plausibly, firms with more commercially valuable IP face greater incentives to redact. This gives rise to the possibility that redaction itself conveys information to investors. We argue that such signaling may help address Arrow's disclosure paradox. While investors do not directly observe the commercial value of the IP licensing deal if it is redacted, they may reasonably infer that redacted agreements are, on average, of higher

⁵³IRS Statistics of Income 2013, "Returns of Active Corporations, Table 6". 2013 is the latest year for which this data is available.

⁵⁴We discuss the definition of materiality and its corresponding threshold in a later section.

value since firms have greater incentives to redact higher value contracts.

We formalize this intuition in a simple, stylized model in which redaction serves as a credible signal of a firm's technology type while shrouding its activities from competitors. A firm faces competition from a potential entrant. If the firm reveals its technology, the potential entrant enters and competes over rents. If the firm redacts, it is able to conceal information contained in its licensing agreement. As a result, the potential entrant is unable to enter the technology or product market space and compete with the incumbent. At the margin, since their technology is more valuable, high-type firms enjoy a greater benefit from redaction than low-types. Therefore, in equilibrium, only high-type firms redact, thus, redaction is viewed as a positive, credible signal by investors. This result gives rise to the first key implication of our model. While redaction mechanically reduces the amount of information available to investors, the fact that a firm redacts itself contains favorable information on the value of the firm's technology, resulting in a positive capital market reaction.

The second key implication of our model is that by preempting potential rivals, firms preserve the economic rents associated with their IP and maintain their competitive advantage. Many licensing agreements cover inherently un-protectable IP, such as trade secrets, or as-yet unpatented inventions that therefore must remain proprietary in nature. Even in a strong IP protection regime, disclosing a licensed technology is problematic because it encourages competitors with substitutive technology to enter the same market. Finally, to the extent that firms' present IP licensing activity is indicative of the nature and scope of their pipeline of impending innovation activities, shielding information on the technology space and economic value of their existing innovation prevents competitors from preempting the firm's impending inventions with their own substitutes. The protection afforded by redaction thus preserves rents for the filing firm, providing a greater incentive to innovate. We thus predict increased innovation and R&D activity for firms that suppress information from competitors.

We empirically test the two implications of our model in a novel database of IP licensing agreements, sourced from material disclosures in SEC filings. The data cover royalties related to patents,

trade secrets, trademarks, copyrights, mining rights and other types of IP. About a quarter of our sample consist of initially redacted agreements, which are uncovered a decade later through Freedom of Information Act requests. At the level of a single disclosure (i.e. an SEC filing), we study the capital market consequences of firms' strategic decision to suppress information about the economic value of their IP as well as their future innovative activity, relative to firms that chose not to redact.

The results of our empirical analysis support the signaling value of strategic redaction as implied by our model. We find that firms which redact enjoy higher stock liquidity post-filing over the short-, mid-, and long-term horizon (two weeks out from the filing date, 1 quarter out and 1 year out), relative to firms which fully disclose. That is, we show that the market rewards, not punishes, firms for non-disclosure. The magnitude is economically large – in the 1 year post-filing, redactors enjoy an increase in liquidity (gauged by Amihud (2002)'s illiquidity measure) that represents 7% of the unconditional mean, relative to non-redacting firms. Complementing this result, redaction also predicts increased equity issuance and institutional ownership. The evidence favors our first key implication that investors seem to recognize and reward strategic redaction of IP licensing agreements.

In terms of the redacting firms' innovation activity, we find that investors' responses are evidently rational in that redaction predicts higher levels of future innovation and innovative activities as early as 1 year out and for several years subsequently. Here, the magnitude is strikingly large – in the 1 year post-filing, redactors produce a greater total-dollar-value of innovation (implied by stock market reactions to patent grants, as in Kogan et al. (2016)), on the order of 50% of the unconditional mean, relative to non-redactors. This suggests that the ability to withhold commercially sensitive information about their IP portfolio enhances firms' ex-ante incentives to innovate, verifying our second key implication.

Of course, in the limit, for signaling to be credible, not all firms, at all times, should suppress information. Costless redaction would enable firms with lower-quality agreements to pool together with better firms, generating an adverse selection problem. We conjecture that the positive capital

market response to redaction is largely reversed when investors seem to perceive higher information asymmetry. This gives rise to the prediction that there should be cross-sectional heterogeneity in the relationship between redaction and liquidity. Indeed, consistent with the idea that redaction is not a credible signal for all firms, we find that firms with lower reputation or higher information asymmetry face a negative marginal relation between redaction and liquidity. These include firms that are small and young, and firms that are repeated redactors, suggesting that the market rewards firms that censor – but only up to a point. Finally, we find a positive marginal effect for firms that are equity-financing dependent, who have to turn to equity markets more frequently to raise capital, and thus are disciplined by this repeated game with investors.

We emphasize three contributions of this chapter. First, we contribute to the literature on the financing of innovation, by analyzing the effect of relaxing a key informational friction (i.e. the disclosure paradox) on the ability of innovative firms to raise financing. In their excellent review, Kerr and Nanda (2015) note four key frictions to financing innovation. First, the payoffs to innovation are uniquely uncertain. Second, the returns to successful innovation are highly skewed, making it hard for potential financiers to accurately value projects and diversify. Third, with heightened information asymmetry, the inability to write state-contingent dynamic contracts is particularly acute for innovative firms. Finally, since the nature of knowledge and IP is non-rival, the disclosures required in the process of raising finance can subject innovators to risk from competitors. The focus of this chapter is the last friction, which has received considerably less attention. While the disclosure issue has also been incorporated into recent theoretical work on the optimal financing of R&D intensive firms (Thakor and Lo, 2016), to our knowledge, ours is the first study to empirically study the disclosure problem directly in the context of innovative firms. Studying the disclosure friction in an empirical setting is difficult, as it is usually impossible to observe the counterfactual of what was not disclosed. However, as our setting relies on redaction of otherwise mandatory corporate filings, we circumvent this problem as we are able to retrieve the originally hidden information long after the fact. Using this setting, we argue that the ability to strategically withhold information on the value of an idea effectively mitigates this friction.

Second, our study speaks to theoretical and empirical work in pursuit of the notion that a well-functioning marketplace for ideas is important in promoting entrepreneurship and innovation (Akcigit, Celik, and Greenwood, 2016; Gans, Stern, and Wu, 2016). Building upon prior empirical studies on trade in the market for ideas through licensing (Gans, Hsu, and Stern, 2008; Hegde, 2014; Hegde and Luo, 2017), we introduce an extensive database on IP licensing. As far as we know, it is the largest and most detailed of its kind to be used in academic research. Amidst growing calls by academics and policymakers for increased transparency and disclosure, our findings should give pause for thought.

Finally, we provide new insights on the costs and benefits of corporate disclosure, particularly in the context of innovative firms. Prior empirical studies on redacted disclosure (Boone, Floros, and Johnson, 2016; Verrecchia and Weber, 2006) find a negative effect of redaction on capital market outcomes, which they argue is evidence that redaction increases information asymmetry problems between firms and investors. Our results, which emphasize a positive capital market reaction, differ due to our focus on innovative firms (i.e. firms that engage in IP licensing⁵⁵), for whom redaction itself is informative, as it signals the possession of a good and costly to disclose idea. Further, Verrecchia and Weber (2006) consider a sample of small firms (between \$50 and \$100 million market capitalization), and Boone, Floros, and Johnson (2016) consider a sample of IPO firms, which are by definition young. These are groups for whom information asymmetry problems are likely to be more acute, potentially outweighing the positive signaling effect. Indeed, even in our sample we find that marginal effect of redaction on stock liquidity is negative for smaller and younger firms.⁵⁶

The remainder of the chapter is organized as follows. In Section 3.2, we describe the institutional context and develop our hypotheses based upon a stylized model of redaction. In Section 3.4, we

⁵⁵The SEC also permits redactions on other types of filings including credit/leasing, customer-supplier, employment, and equity related agreements. Our sample consists exclusively of licensing agreements, whereas the samples of Boone, Floros, and Johnson (2016) and Verrecchia and Weber (2006) consider redactions on all types of filings. Our sample thus allows us to study the role of redaction of specific, IP-related information, corresponding most closely to the tradeoff implied by Arrow's disclosure paradox.

⁵⁶Our findings on the benefits of non-disclosure for innovative firms complement the SEC's existing concerns that more disclosure is not better in the sense it creates an information overload for investors. For more details, see <https://www.sec.gov/News/Speech/Detail/Speech/1365171492408>, retrieved 3/11/2017.

describe our data sources and provide descriptive statistics. In Section 3.6 we present our empirical analysis on the consequences of redaction and in Section 3.7, we conclude.

3.2 Institutional Context and Hypothesis Development

3.2.1 Institutional Context

Since as far back as the Securities Act of 1933, companies with publicly sold securities are required to state all material facts relevant to investors who might buy their securities. As a guiding principle, firms are required to disclose “material definitive agreements not made in the ordinary course of business”. The definition of materiality was sharpened through a variety of historical developments. The modern standard of materiality was refined in the 1976 case *TSC Industries, Inc. v Northway Inc.*, which defined materiality as a fact which, if omitted, would have significantly altered the “total mix of information available” in a shareholder’s decision to purchase a security or vote.⁵⁷

In spite of broad guidance provided by regulatory agencies, the decision on whether a given corporate development crosses the materiality threshold is often subject to the discretion of the company. Some rules-of-thumb exist. For example, the 5% rule of thumb suggests that reasonable investors would not be influenced in their investment decisions by actions that lead to fluctuations of net income of 5% or less.⁵⁸ For a detailed discussion on the considerations involved in auditors’ determination of materiality, see Choudhary, Merkley, and Schipper (2017). Regardless of the exact threshold, material disclosures represent significant events, the knowledge of which affects the decisions of current and prospective investors.

However, recognizing the potential commercial harm of disclosing proprietary information, the

⁵⁷In the words of Justice Thurgood Marshall, “An omitted fact is material if there is a substantial likelihood that a reasonable shareholder would consider it important in deciding how to vote. . . . It does not require proof of a substantial likelihood that disclosure of the omitted fact would have caused the reasonable investor to change his vote. What the standard does contemplate is a showing of a substantial likelihood that, under all the circumstances, the omitted fact would have assumed actual significance in the deliberations of the reasonable shareholder. Put another way, there must be a substantial likelihood that the disclosure of the omitted fact would have been viewed by the reasonable investor as having significantly altered the ‘total mix’ of information available.”

⁵⁸See the Journal of Accountancy, <http://www.journalofaccountancy.com/issues/2005/may/thenewimportanceofmateriality.html> retrieved 12/10/2016.

SEC has allowed for firms to request confidential treatment of certain information in material filings, as stipulated in Rule 406 under the Securities Act of 1933 and Rule 24b-2 under the Securities Exchange Act of 1934. The Staff Legal Bulletin No. 1 issued in February 1997 and its Addendum issued in July 2001,⁵⁹ on the subject of “Confidential Treatment Requests” (CTR) by the Division of Corporation Finance, contain guidance for filing firms requesting confidential treatment. From this, we are able to infer details about the SEC’s interpretation of confidential treatment requests. In particular, the guidance explicitly notes the proprietary cost argument for requesting redaction:

“Sometimes disclosure of information required by the regulations can adversely affect a company’s business and financial condition because of the competitive harm that could result from the disclosure. This issue frequently arises in connection with the requirement that a registrant file publicly all contracts material to its business other than those it enters into in the ordinary course of business. Typical examples of the information that raises this concern include pricing terms, technical specifications and milestone payments.”

The guidance specifies certain requirements for CTRs. First, they should not be overly broad. Specifically:

“The information covered by an application should include no more text than necessary to prevent competitive harm to the issuer.”

Second, applicants must provide written justification for the exemption. Third, applicants must specify a particular duration. This portion was later clarified in the Amendment to provide specific guidance:

“If the remaining term of the contract is greater than 10 years from the date of the extension application, we generally will only grant confidential treatment for 10 years; If the remaining term of the contract is less than 10 years from the date of the extension application, we will consider a request for the remaining term of the contract; If the

⁵⁹Full text is available at <https://www.sec.gov/interp/legalslbcf1r.htm>, retrieved 3/11/2017.

remaining term of the contract is less than five years from the date of the application, but there is a possibility that it will be extended beyond its stated term, we will consider granting confidential treatment for a period of up to five years.”

Fourth, applicants must identify clearly the information that is the subject of the application. Finally, applicants must consent to the release of the information to the SEC for official purposes, for instance evaluation or audit.

In Figure 3.1, we give an example of a portion of a redacted licensing agreement between Genome Therapeutics Corporation and Biomerieux Incorporated dated September 30, 1999. In the section of the agreement that we have reproduced in Figure 3.1, payment terms including royalty rates, sub-licensing fees etc. are redacted.

Under the Freedom of Information Act (FOIA), the public can request the entire uncensored contents of specific filings after the expiration of the CTR. The dataset used in this chapter is partly assembled through thousands of FOIA requests issued to the SEC (in most cases) 10 years after the date of the original filing. The remainder of the dataset consists of filings that were unredacted as of the time of dissemination. As we can compare information that was disclosed with contemporary, counterfactual information that was withheld from the public eye, we have an ideal setting in which to test the impact of non-disclosure.

3.3 Model and Hypothesis Development

In order to motivate our empirical hypotheses, we present a simple, stylized model⁶⁰ that demonstrates the strategic role of redaction as a signal of technology quality. The setup consists of three firms, an innovative firm (whose choice of redaction we are modeling), an uninformed rival whose

⁶⁰Our model follows in the spirit of several theoretical models of voluntary disclosure. The theoretical literature, starting with Verrecchia (1983), models the tradeoff faced by firms making the decision to voluntarily disclose information to investors, which is presumed to improve the precision with which the market can value their stock, but causes them to incur proprietary costs. Verrecchia (1983) establishes the existence of a partial disclosure equilibrium, and shows that the greater the proprietary cost of disclosure, the less negative traders’ reaction to the withholding of information. Further studies have endogenized the proprietary cost of disclosure as a function of product market rivals’ responses to the disclosed information (Bhattacharya and Ritter, 1983; Darrough and Stoughton, 1990; Feltham and Xie, 1992; Wagenhofer, 1990).

6.5 ROYALTIES PAYABLE BY EACH PARTY AND ITS AFFILIATES AND SUBLICENSEES.

6.5.1 ROYALTIES ON NET SALES OF BMI PRODUCTS AND GTC PRODUCTS.

- (a) BMI shall pay royalties to GTC based on the Net Sales of a BMI Product at the rate, *...*, specified in EXHIBIT G hereto. *...*.

In the event a BMI Product should subsequently become a BMI Blockbuster Product, the Parties shall determine, within *...* days following the BMI Product's change of status, a new royalty rate to be applied to the BMI Blockbuster Product, it being understood that such rate shall be *...*.

(b) GTC, its Affiliates and sublicensees shall pay royalties to BMI based on the Net Sales of a GTC Product at a rate to be determined in good faith by mutual agreement of the parties with respect to each GTC Product within *...* days following the First Commercial Sale of the GTC Product. The royalty rate for Net Sales of each GTC Product shall be determined *...*.

6.5.2 SUBLICENSES. (i) BMI and GTC shall, with respect to any sublicensees by BMI or GTC hereunder, pay to the other party *...* percent (*...*) of all consideration received

For the purposes of calculating the royalty payments to be made in United States dollars, GTC shall multiply the sales made in countries other than the United States by the appropriate royalty rate before converting such sum into United States dollars at the exchange rate used by GTC for reporting such sales for United States financial statement purposes.

BMI shall use, for the purposes of calculating the royalty payments to be made in United States dollars, the following rules:

- (a) For the period from January 1, 1999 until *...*:

For the conversion of all currencies, other than United States dollars, into United States dollars, the amount shall be first converted into the Euro using the official spot rate published by the European Central Bank on the last business day of the period to which the payment of royalties relates and then, the amount so obtained in the Euro shall be converted into United States dollars applying the official rate published by the European Central Bank on the last business day of the period to which the payment of royalties relates.

- (b) For the period from *...* until the expiration of the contract:

(i) For the conversion of the Euro into United States dollars, the amount shall be converted applying the official rate published by the European Central Bank on the last business day of the period to which the payment of royalties relates.

(ii) For the conversion of currencies other than the Euro into United States dollars, the amount shall be first converted into the Euro using the official spot rate published by the European Central Bank on the last business day of the period to which the payment of royalties relates and then, the amount so obtained in the Euro shall be converted into

Figure 3.1: Sample Redacted Licensing Agreement

This figure is an example of a portion of a redacted licensing agreement between Genome Therapeutics Corporation and Biomerieux Incorporated dated September 30, 1999. In the section of the agreement that we reproduced in this figure (the total length of such agreements often exceed 30 pages), payment terms including royalty rates, sub-licensing fees etc. are redacted.

decision to enter the innovative firm's technology space and compete is a function of the innova-

tive firm's redaction decision, and an investor, whose preferences and beliefs determine the market price of the innovative firm's stock. The innovative firm, which we denote as A , is endowed with a technology type $t \in \{\bar{t}, \underline{t}\}$, with $\bar{t} > \underline{t}$, where the unconditional probability $\mathbb{P}(t = \bar{t}) = \theta$ is known by the uninformed potential rival, which we denote as B , and the investor, whom we denote as I . However, the realization of the technology type, \tilde{t} , is known only to A .

Firm A has the choice to disclose \tilde{t} , and we denote this choice by $D \in \{0, 1\}$. The choice by firm A to disclose (which is exactly equivalent to the choice to not redact), $D = 1$, results in both B and I observing the reported value of t , \hat{t} . In order to narrow the strategy space, we assume that firm A 's disclosure of \tilde{t} is truthful, i.e. the reported $\hat{t} = \tilde{t}$. Thus, for the sake of convenience we simply use \tilde{t} to denote firm A 's disclosure. The assumption of truthful reporting is reasonable given that the disclosures in question are SEC filings, subject to auditor approvals as well as SEC review through the Division of Corporation Finance, and costly shareholder litigation if proven to be false. The choice by A to not disclose (which is exactly equivalent to the choice to redact), $D = 0$, means that both B and I are unable to observe \tilde{t} . In case of complete non-disclosure regardless of A 's type, they both believe that \tilde{t} takes its unconditional expected value $t' = \theta\bar{t} + (1 - \theta)\underline{t}$.

In addition, firm A is associated with an ex-ante degree of information asymmetry, $\beta \in \{\beta_G, \beta_B\}$ where $\beta_G < \beta_B$. Firm A has $\beta = \beta_G$ with probability λ , and this is independent of its technology type t . Let the unconditional expected value of β be defined as $\beta' = \lambda\beta_G + (1 - \lambda)\beta_B$.

Our analysis proceeds in two steps. First, we consider the effect of A 's disclosure decision on B 's entry decision, and the resulting effect on A 's profits. Second, we analyze the investor's response to the product-market equilibrium. For the first step, we assume that the economic rents available in the market occupied by A , which B is considering entering, are a function of A 's technology type \tilde{t} . Specifically, we assume that the rents are normally distributed with mean \tilde{t} and variance $\frac{\sigma^2}{\tilde{t}}$, i.e. we denote the total rents available in the market as π and $\pi \sim N\left(\tilde{t}, \frac{\sigma^2}{\tilde{t}}\right)$. The rents (net of costs) earned by each firm are indicated by π^A and π^B .

We now consider Firm B . Firm B incurs a cost $c > 0$ of entry, if it chooses to enter the market. B also incurs s , a cost of search, if it wishes to enter the market when firm A does not disclose its

type, i.e. $D = 0$. The search cost s is a function of B 's belief of A 's type, which we denote t^* . We do not specify a functional form for s , only that $s(\underline{t}) = \underline{s}$ and $s(\bar{t}) = \bar{s}$, with $\bar{s} > \underline{s} > 0$, and $\frac{\partial s}{\partial t^*} > 0$. The search cost may, for example, correspond to the fact that the company B must conduct research to determine the nature of the undisclosed technology so as to imitate it. Also, it must develop its own substitutive technology whose cost is increasing in the type of technology - that is, it is harder to replicate a good technology than a bad technology.

How does Firm B 's decision affect payoffs for A ? Since A is already in the industry, it does not incur any cost. Both firms A and B are assumed to be risk-neutral profit maximizers. If B does not enter, we assume that the entire rents are received by A , that is $\pi^A = \pi$. If B enters, the two firms split the rents evenly. So, if B enters, A receives $\pi^A = \frac{\pi}{2}$. For B , if it enters when A has selected $D = 1$ (has chosen to disclose), its expected profit is $\mathbb{E}[\pi^B] = \mathbb{E}\left[\frac{\pi}{2}\right] - c$. If B enters when A has selected $D = 0$, its expected profit is $\mathbb{E}[\pi^B] = \mathbb{E}\left[\frac{\pi}{2}\right] - c - s$. Firm B only enters if, based on its beliefs on \tilde{t} , $\mathbb{E}[\pi^B] \geq 0$, namely $\mathbb{E}\left[\frac{\pi}{2}\right] \geq c + s$. We further assume that $\frac{t}{2} < c$, implying that when B knows A 's type is low ($\tilde{t} = \underline{t}$), it does not enter as $\mathbb{E}\left[\frac{\pi}{2}\right] = \frac{t}{2} - c - s < 0$. We also assume that $\frac{\bar{t}}{2} < \bar{s}$, indicating that the search cost incurred by B when A 's technology is high-type is prohibitive.

To ascertain the equilibrium outcome, consider Firm A 's strategies, which can be represented by pairs $\{D(\underline{t}), D(\bar{t})\}$. For example, $\{1, 1\}$ implies that firm A discloses its type both when $\tilde{t} = \underline{t}$ and $\tilde{t} = \bar{t}$. In Appendix C.1, we evaluate four cases, corresponding to the four strategies available to firm A (as it has two possible actions and can be of two possible types). This analysis shows that as long as the probability θ exceeds a lower-bound value (i.e. there exists sufficiently large probability that A 's technology is high-type), the game has two Nash equilibria in pure strategies. By iterated elimination of weakly dominated strategies the unique surviving Nash equilibrium strategy is $\{1, 0\}$, where A discloses if its technology is low-type and redacts if its technology is high-type (see Appendix C.1 for details). This is an essential condition for non-disclosure (redaction) to be a credible signal of type. The equilibrium is separating, and as in Gertner, Gibbons, and Scharfstein (1988), it is the nature of the product-market competition that determines whether the equilibrium is separating or pooling.

We now turn to the second part of our analysis, determining the investor's response to the product-market equilibrium. In equilibrium, it is optimal for A to redact only if $\tilde{t} = \bar{t}$ and disclose only if $\tilde{t} = \underline{t}$. The investor thus infers that $\tilde{t} = \bar{t}$ if she observes non-disclosure and directly observes that $\tilde{t} = \underline{t}$ when the technology is disclosed. The equilibrium payoff to firm A from its optimal strategy of $\{1, 0\}$ is $\{\underline{t}, \bar{t}\}$. We assume that the world ends after the product market competition takes place, and the investor receives firm A 's total profit π^A as a liquidating dividend. As in Verrecchia (1983), the price of the firm's stock is determined by the following expression reflecting the investor's mean-variance preferences:

$$P = \mathbb{E}[\pi^A] - \beta \text{Var}[\pi^A]$$

Since the equilibrium is separating, the investor prices A 's stock as:

$$P = \begin{cases} \bar{t} - \beta_G \frac{\sigma^2}{\bar{t}} & \text{if } D = 0 \text{ and } \beta = \beta_G \\ \bar{t} - \beta_B \frac{\sigma^2}{\bar{t}} & \text{if } D = 0 \text{ and } \beta = \beta_B \\ \underline{t} - \beta_G \frac{\sigma^2}{\underline{t}} & \text{if } D = 1 \text{ and } \beta = \beta_G \\ \underline{t} - \beta_B \frac{\sigma^2}{\underline{t}} & \text{if } D = 1 \text{ and } \beta = \beta_B \end{cases}$$

The expected price conditional on observing redaction, i.e. $D = 0$, is $\mathbb{E}[P|D = 0] = [\lambda \times (\bar{t} - \beta_G \frac{\sigma^2}{\bar{t}})] + [(1 - \lambda) \times (\bar{t} - \beta_B \frac{\sigma^2}{\bar{t}})] = \bar{t} - \beta' \frac{\sigma^2}{\bar{t}}$, and the expected price conditional on observing $D = 1$, i.e. full disclosure, is $\mathbb{E}[P|D = 1] = \underline{t} - \beta' \frac{\sigma^2}{\underline{t}}$. The variance of the price conditional on observing $D = 0$ is $\text{Var}[P|D = 0] = \tilde{\beta} \frac{\sigma^4}{\bar{t}^2}$ and the variance of the price conditional on observing $D = 1$ is $\text{Var}[P|D = 1] = \tilde{\beta} \frac{\sigma^4}{\underline{t}^2}$, where $\tilde{\beta} = \lambda(1 - \lambda)(\beta_G - \beta_B)^2$. As $\tilde{\beta} > 0$ and $\bar{t} > \underline{t}$, this implies that $\text{Var}[P|D = 0] < \text{Var}[P|D = 1]$.

Our model generates two main implications, and one cross-sectional implication that we take to the data.

Implication 1 The first key implication of our stylized model is that redacting firms have a lower conditional variance of stock price as $\text{Var}[P|D = 0] < \text{Var}[P|D = 1]$. To the extent that

the conditional variance of a stock's price is a proxy for the stock's illiquidity, this implies that redacting firms experience better stock liquidity than non-redacting firms.

Implication 2 The second key implication is that redaction is associated with greater innovation, as firms with high-value technology are able to preserve the economic rents from potential competitors and avoid expropriation. This is consistent with the fact that in equilibrium, the rival firm B does not enter A 's market and compete.

Cross-sectional Implication If investors are able to observe β , then following a similar argument that we used to demonstrate that $Var [P|D = 0] < Var [P|D = 1]$, it can be shown that $Var [P|\beta_G] < Var [P|\beta_B]$. Thus, in the cross-section, firms which are perceived as having a worse reputation, or a greater degree of information asymmetry, by investors (i.e. firms with $\beta = \beta_B$) should experience worse stock liquidity conditional on a given disclosure choice.

3.4 Data

3.4.1 Intellectual Property Licensing Data

The main data used in this chapter are drawn from a database of intellectual property licensing agreements from RoyaltyStat LLC. RoyaltyStat is one of a handful of companies that collect data on royalty rates of IP licensing transactions, and was the first company to come to market with a searchable online platform in 2000. The data are sourced from SEC-mandated material disclosures, wherein the licensing agreements are usually attached as exhibits in both scheduled and one-off filings. RoyaltyStat has developed a natural language processing engine to identify relevant filings and automatically populate certain standardized fields. Analysts then review and augment the output. The data are updated daily.

RoyaltyStat provided us access to every record and a rich set of fields in their databases. To preserve data confidentiality, we did not custody the data - we were provided access through a password-protected portal in which we uploaded R scripts and received textual output. Table 3.1 enumerates some of the key fields we derived from their data. We use these fields to account for

Table 3.1: Data Fields Extracted from RoyaltyStat Data

| Identifying Information | Description |
|-------------------------------------|--|
| Licensor (name, CIK, <i>gvkey</i>) | The party in possession of the intangible asset contracting to allow another party to use it in exchange for payment. |
| Licensee (name, CIK, <i>gvkey</i>) | The party paying the possessor of the intangible asset in exchange for the right to use the intangible for proprietary purposes. |
| Filer (name, CIK, <i>gvkey</i>) | The party which submitted a filing to the SEC documenting the existence of a contract. The filer is usually (but not always) identifiably the licensor or licensee. |
| <i>IsRelatedParty</i> | Are the parties subsidiaries of the same company? For the purpose of transfer pricing, this distinction is made because the contract does not represent open-market pricing. |
| Filing Information | |
| Filing date | The date of the filing in EDGAR. |
| FOIA | Was the record originally redacted, and later recovered, through the Freedom of Information Act? |
| Intangible information | |
| SIC of intangible | Based on the technology described in the SEC filing, RoyaltyStat's analysts assign the technology covered in a contract a Standard Industrial Classification code. |
| RoyaltyStat Industry | RoyaltyStat has a classification of 42 industries. Possible values include: Pharmaceutical Biotech, Medical Devices, and Software. |
| RoyaltyStat Subindustry | Each industry is classified into a number of sub-industries. For instance, the Pharmaceutical Biotech industry contains Cancer, Genetic, and Cardiovascular sub-industries. |
| Intangible numbers | Intangible numbers specifically mentioned in a filing. The vast majority are patents, including both applications and grants. |
| Description | A textual description of the contract and intangible being licensed. |
| Intangible property type | RoyaltyStat's proprietary categorization of intangibles in a contract. Possible values include patents, trademarks, copy rights, trade secrets, software, process know-how, land and mineral rights. |

| Contract features | |
|-----------------------------|---|
| Duration | The number of years, if known, a royalty contract is active for. In some contracts, the duration is unknown, perpetual, or is stated but the contract contains an option to renew. In econometric analysis, if it is perpetual, it is coded as 100 years. |
| Agreement type | Whether the agreement can be classified as a cross-license, joint-venture, patent-related, trademark-related, and so forth. From this we create booleans for each agreement type. |
| Rights granted | For instance, whether the contract gives the licensee the right to sublicense |
| Territory agreement applies | The countries or legal jurisdictions over which the company is granted a license to use the technology. Possible values include specific countries, worldwide, cyberspace, or unknown. |
| Effective date | The date the contract terms become effective, if known. In some cases, the effective date was immediate and the filing date occurred with a delay, and in some cases the filing date precedes the effective date |
| Exclusivity | Is the contract exclusive to the party? |
| Payment information | |
| Royalty base | The accounting identity against which any royalty rate is assessed. Most commonly it is net sales. Others include cost-of-goods sold, net profits, per-yard, per square meter, per unit. |
| Royalty rate | In the units of the royalty base, this is the numerical payment made by the licensee. |
| Royalty rate currency | The currency against which per-unit royalties are assessed. |
| Tiered royalty | Many royalty contracts have incentive schemes whereby the rate changes at variable quantities. |
| Royalty rate low | When the rate is tiered, this is the lowest rate that can be charged, if stated. |
| License fee amount | The currency amount of the license fee. |
| License fee currency | The currency in which the license fee is to be paid. |
| Fee type | The vast majority of cases are upfront fees, but in some cases there are scheduled fees, or renewal fees. |
| Other payments | Booleans that flag additional payment terms (upfront fees, renewal fees, milestone payments, etc). |

contract characteristics in our analysis.

The unit of observation is a licensing contract entered into by two entities, a licensor and a licensee. The filer must be a public firm, and is usually either the licensor or the licensee. The counterparty may not be a commercial entity (for example, a non-profit hospital, university, or individual inventor), or may be a private or publicly-listed firm. In addition to identifying information, we observe when the filing was made, and whether or not the record obtained was retrieved through FOIA (i.e. its contents were redacted at the time of filing).

In terms of the contract itself, we observe contract terms, payment terms and intangible information. Contract information consists of the effective date of the contract, the duration, the geographic scope of the licensing agreement, and the the rights granted by the licensing firm to the licensee. Payment terms consist of royalty rates, any upfront fees, and any additional payment terms such as tiered royalty rates that change as the quantity sold increases, or milestone payments for reaching certain contracted goals. Finally, intangible information includes descriptors of the technology, its applicable industry, and where relevant, intangible numbers such as patent grant or application numbers.

Using this data, we then construct a sample. The details of the sample selection are presented in Appendix C.2. RoyaltyStat maintains a link to Compustat as a sub-licensor of Compustat data. We assume their link is valid if available - if missing, we augment their linkage using either the *CIK* or filer name using various linktables. Table 3.2 reports the sample sizes at each stage of the filtering process. We have 7,972 license agreements whose filer can be linked to Compustat and can be disambiguated as either the licensor or licensee, and of them, 6,579 are US firms with publicly-traded equity covered by CRSP at the time of filing. Some firms which do not make it into our sample are penny stocks not covered by Compustat, unlisted firms, or even Canadian or Mexican public firms. Our regressions also lose observations due to missing values for control variables such as assets, leverage, or various contract characteristics. However, our qualitative inferences are not sensitive to excluding either contract or firm characteristics. Finally, we also lose observations due to missing data for outcome variables.

In most specifications, we have between 3,000 to 4,000 observations. With this sample size, we make the claim this is the first large-sample analysis of licensing agreements to our knowledge, as it is larger than the handful of studies with access to licensing contract-level data. Our sample is about as large as a typical study on syndicated loan contracts or mergers.⁶¹

⁶¹We have a similar sample size to, for instance, Campello and Gao (2016).

Table 3.2: Sample Selection Procedure

| | # | Comment |
|---|-------|--|
| # Observations to start. | 17487 | As of June 2016, this was the entire number of records available on RoyaltyStat's database. |
| # Obs with filer <i>gvkey</i> matched to Compustat North America. | 8774 | Based on RoyaltyStat's maintained <i>gvkeys</i> , we can identify the following number of firms. |
| # Obs of US contracts with <i>gvkey</i> after programmatically imputing using <i>CIK</i> or name. | 9678 | We augment RoyaltyStat's <i>gvkeys</i> using programmatic matching of <i>CIK</i> codes to <i>gvkeys</i> based on Capital IQ Helper. We also match based on names. If a punctuation-cleaned name is an exact match for a Capital IQ Company name, CRSP name, Compustat name, Worldscope name, which can be linked back to a <i>gvkey</i> , it is in our database. |
| # Obs where the identity of the filer can be disambiguated (i.e. the filer can be assigned the status of being either the licensor or the licensee), excluding related party filings. | 7972 | This is our baseline sample. |
| # Obs of US contracts with <i>PERMNOs</i> . | 6579 | Not all companies have <i>PERMNOs</i> because they are not exchange-listed, the observation is pre-listing or post de-listing. |

3.4.2 Other Data

We merge the Compustat sample with several other databases. We aim to capture two main categories of outcomes: capital market outcomes and innovation related outcomes. Many of the variables we use limit our sample size due to time period coverage, especially the patent data.

Capital Market Outcomes We obtain data on capital market consequences from CRSP, DTAQ, SDC Platinum and Thomson Reuters 13/F. From the CRSP Daily Stock File, we calculate the daily bid-ask spread, defined as the closing ask minus the closing bid scaled by the end-of day price, and the daily turnover, defined as end-of-day price multiplied by volume of shares traded. For every contract i filed by firm j on date t , we average these daily measures in the subsequent 1 year. We also construct the stock illiquidity measure of Amihud (2002), by computing in each day

the average ratio of the daily absolute return to the (dollar) trading volume on that day. This daily measure, too, is averaged into an annual measure over the subsequent 1 year for every contract i filed by firm j on date t .

From the NYSE Daily Trades And Quotes database, we calculate the price impact of trades and effective spread. The price impact of trades, also known as Kyle's Lambda, is measured by performing regressions of the absolute return over a given time period on dollar volume. The price impact is defined as the slope from this regression. These slopes are averaged to calculate daily and annual measures of price impact in an analogous way to the spread and turnover obtained from CRSP. Higher values of price impact correspond to a given magnitude of trading being associated with a greater movement in returns, and thus price impact is seen as a measure of stock illiquidity (in fact, this is an intra-day high frequency analogue of the (Amihud, 2002) measure which we compute from CRSP). The weekly effective spread is defined as twice the absolute value of the difference between the actual trade price and the midpoint of the market quote (i.e., between the quoted bid price and the quoted ask price), divided by the midpoint between these two prices based on DTAQ.

We use the SDC Platinum database in order to construct a dummy variable $\mathbf{1}\{Issuance\}$ that takes the value of 1 if a given firm conducts a secondary equity offering (SEO), in which capital is raised through an equity issuance, in a given year. The Thomson Reuters 13/F database is used to calculate institutional ownership (percentage of shares outstanding held by 13/F filing institutions), ownership by large owners (percentage of shares outstanding held by the top 5 largest institutional owners), and concentration (Herfindahl-Hirschman Index of institutional ownership by 13/F filing institutions).

Innovation Outcomes The main innovation outcomes we consider are related to patenting. Although patenting is not the only type of innovation, nor the only type of innovation that firms license, it is the easiest to objectively measure and quantify. About 54% of contracts in our sample are patent-related (with 1/4th having intangibles information related to patents). For the contracts with intangibles that refer to *granted patents*, we use the Berkeley patent database to calculate nu-

merous patent statistics related to the number of citations, the number of claims made in the patent, and the features of the bundle as a whole. To describe innovation activity at the firm level, we use the NBER 2006 patent sample, which covers all firms in North America and accounts for mergers and delistings, through 2006. For just the US-listed sample, we also use the patent data provided by Kogan et al. (2016) through 2010 to obtain firm-level innovation measures. Specifically, we consider the firm-level annual measures of patenting, *TCW* and *TSM*. *TCW* refers to the forward citation weighted count of all patents granted to a given firm in a given year, with adjustments for citation truncation lags. *TSM* refers to the total dollar value of patents granted to a given firm in a given year, estimated from the stock market reaction to the patent grants. Detailed definition of both variables are contained in Kogan et al. (2016). We extend the patents outward through 2014 using the Berkeley patent data to correct for the truncation issue raised in Lerner and Seru (2015).

3.4.3 Summary Statistics

In this section, we provide basic summary statistics to familiarize the reader with our dataset, and, more broadly, the world of intellectual property licensing. In Figure 3.2 we plot the distribution of contracts across industries, where the industry classification scheme used is the 42 technology industries defined by RoyaltyStat. Specifically, each contract is classified into 1 (or more) of the 42 technology industries based upon the technology space that best describes the underlying IP being licensed (this determination is made by RoyaltyStat analysts). As would be expected in a study focusing on IP-intensive firms, the vast majority of the licensing contracts have underlying IP that falls under the categories of Pharmaceutical Biotech, Medical Devices and Software.

In Figure 3.3 we plot the yearly distribution of total contracts filed and the proportion which are redacted. Note that because redacted contracts only enter into the RoyaltyStat database once their analysts are able to make an FOIA request to retrieve the original un-redacted filing, this implies that the CTR status of these contracts must have expired. As discussed in the Section 3.2.1, a CTR lasts around 10 years on average. Therefore, post-2006 we are significantly undersampling redacted contracts. This is simply because for the average redacted contract filed post-2006, the

Table 3.3: Contract-level Summary Statistics

| Panel A: Distribution of Contracts by Royalty Base | | | |
|--|-----------|------------|--|
| Royalty Base | Frequency | % of total | |
| Cost of Goods Sold | 47 | 0.60% | |
| Gross Profit | 297 | 3.70% | |
| Manufacturing Cost | 34 | 0.40% | |
| Net Asset Value | 15 | 0.20% | |
| Net Profit | 230 | 2.90% | |
| Net Sales | 6704 | 84.10% | |
| Net Smelter Returns | 411 | 5.20% | |
| Operating Profit | 139 | 1.70% | |
| Overriding Royalty | 93 | 1.20% | |
| Value Added | 2 | 0.00% | |
| Total | 7972 | 100% | |

| Panel B: Distribution of Key Contract Terms | | | |
|---|------------------|------------------|------------------|
| | License Fee (\$) | Royalty Rate (%) | Duration (years) |
| N | 2418 | 7972 | 6625 |
| Mean | 4280652 | 11.49 | 19.744 |
| St. Dev. | 23279623 | 14.363 | 26.049 |
| Min | 0 | 0.005 | 0 |
| Q1 | 2 | 0.5 | 0 |
| Q5 | 5000 | 1 | 1 |
| Q25 | 45792 | 3 | 5 |
| Q40 | 100000 | 5 | 10 |
| Median | 250000 | 6 | 17 |
| Q60 | 500000 | 8 | 20 |
| Q75 | 1500000 | 12 | 20 |
| Q95 | 15000000 | 50 | 99 |
| Q99 | 60000000 | 70 | 99 |
| Max | 550000000 | 100 | 500 |

| Panel C: Distribution of Contracts by IP type | | | |
|---|---------------|----------------|---------------|
| IP Type | Frequency (%) | IP Type | Frequency (%) |
| Is Amendment | 13.40% | Is Patent | 54.00% |
| Is Asset Purchase | 0.00% | Is Process | 14.70% |
| Is Consulting | 0.90% | Is Proprietary | 2.30% |
| Is Copyrights | 15.70% | Is Research | 6.30% |
| Is Cross-License | 2.10% | Is Services | 3.30% |
| Is Distribution | 4.20% | Is Shares | 1.40% |
| Is Franchise | 3.50% | Is Show-How | 16.80% |
| Is Know-How | 30.50% | Is Software | 14.70% |
| Is Liabilities | 2.20% | Is Sublicense | 2.90% |
| Is Marketing | 1.30% | Is Supply | 3.60% |
| Is Mineral | 9.30% | Is Technology | 30.80% |
| Is Option | 4.60% | Is Trademark | 26.00% |

CTR is yet to expire. Thus, the original un-redacted contract cannot be retrieved via an FOIA request.⁶² In the pre-2007 period, the rate of redaction appears to be growing over time, and over the 2000-2006 period in which we observe the most contracts (both redacted and public), the proportion of total contracts that are redacted ranges between 30% and 45%.

The panels of Table 3.3 display summary statistics of various contract characteristics. In panel A, we show the distribution of contracts by royalty base. The vast majority of contracts are based upon net sales. In panel B, we present summary statistics on three key contract features - the license fee, the royalty rate and the contract duration. The summary statistics confirm that these are generally economically significant contracts. The median license fee of \$250,000 represents 2.3% of the median annual sales of a firm in our sample. These contracts are of a relatively long duration, with the median contract having a maturity of 17 years. Also, the distributions of rates and fees are highly skewed, with a long right tail.

In panel C, we show the distribution of agreements by the type of underlying IP. Note that these categories are not mutually exclusive as the IP underlying a single agreement often encompasses several types. Hence, the percentages sum to greater than 100. By far the most common form of IP is patents, followed by know-how, technology, and trademarks.

⁶²Since in the post-2006 period we are significantly undersampling redacted contracts, we rerun our entire analysis in the pre-2007 sample, and aside from a drop in sample size of around 10% to 15%, our results are qualitatively and quantitatively similar.

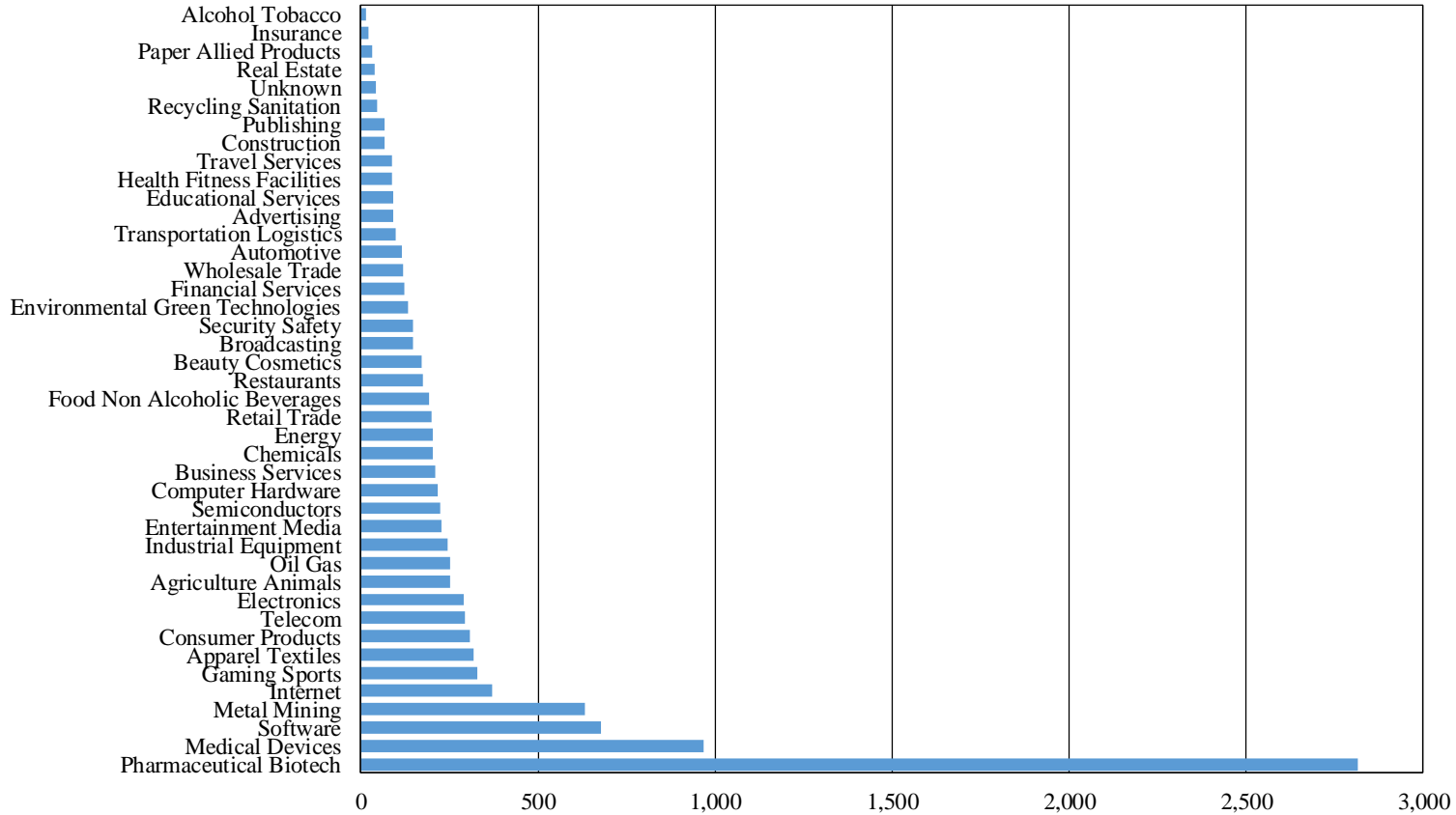


Figure 3.2: Distribution of Contracts by Technology Industry

This figure plots the distribution of contracts across industries, where the industry classification scheme used is the 42 technology industries defined by RoyaltyStat and their analysts. Specifically, each contract is classified into 1 (or more) of the 42 technology industries based upon the technology space that best describes underlying IP being licensed (this determination is made by RoyaltyStat analysts).

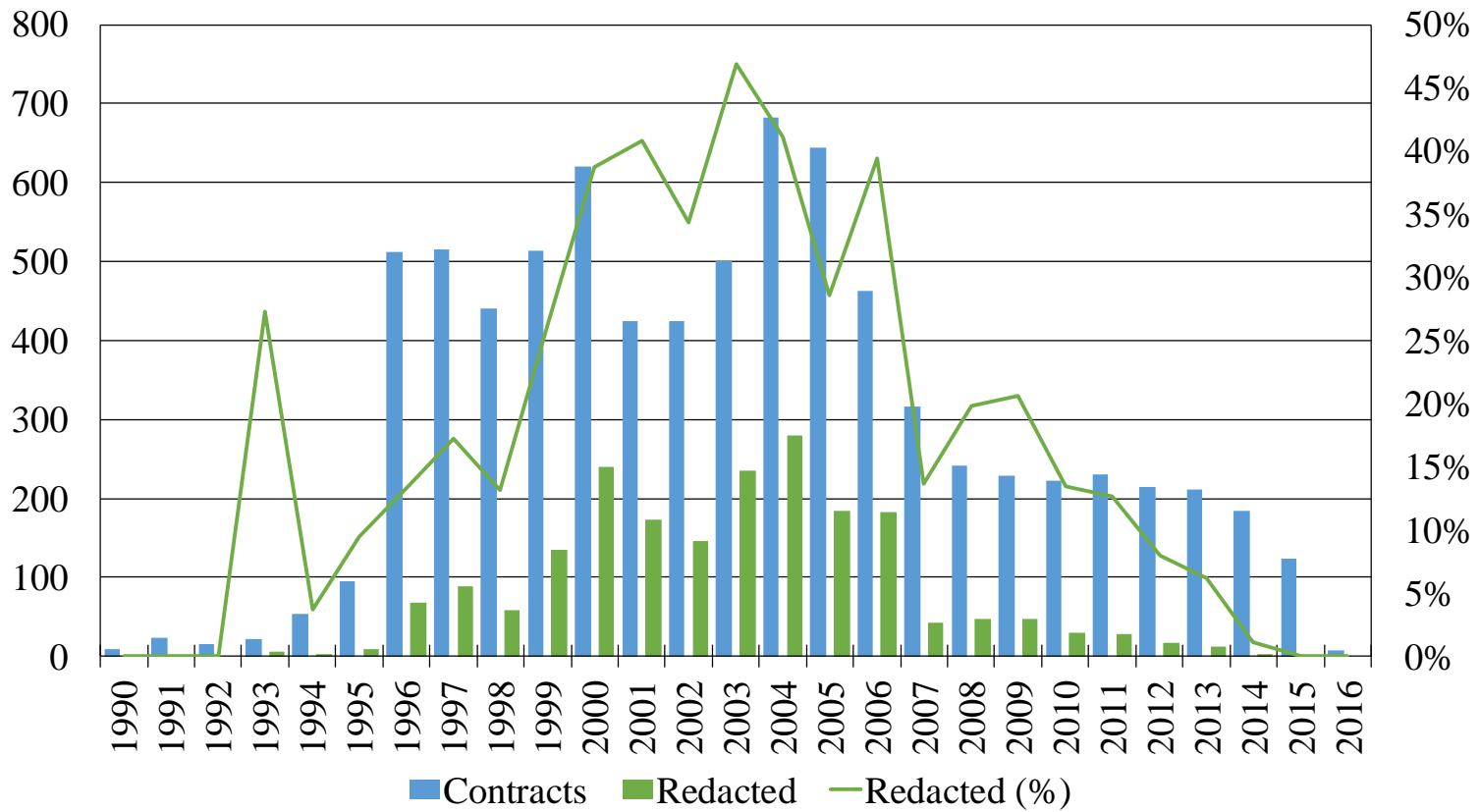


Figure 3.3: Distribution of Redacted vs. Un-redacted Contracts Over Time

This figure plots the yearly (time-series) distribution of total contracts and the proportion which are redacted.

Table 3.4: Firm-Level Summary Statistics

In this table, we compare financial characteristics of the firms which appear in our sample, and the financial characteristics of firms which appear in the overall Compustat sample. Panel A provides summary statistics for key financial variables for the firms in our sample, and Panel B provides summary statistics for the same variables for all firms in the Compustat sample.

| Panel A: RoyaltyStat Sample | | | | | | | | | | | | |
|----------------------------------|-----------|----------|----------|-------------|-----------------------|------------------------|-----------------------|----------------------------|-------------|-----------|-----------|--------------|
| | Tobin's Q | Leverage | Assets | Tangibility | $\frac{R\&D}{Assets}$ | $\frac{Capex}{Assets}$ | $\frac{Cash}{Assets}$ | $\frac{NetIncome}{Assets}$ | Age (Years) | # Patents | Net Sales | Int. Capital |
| N | 5601 | 7281 | 7361 | 7293 | 7361 | 7202 | 7312 | 7320 | 7619 | 4206 | 7320 | 7552 |
| Mean | 2.565 | 0.307 | 343.285 | 0.195 | 0.317 | 0.060 | 0.380 | -0.767 | 12.962 | 4.040 | 202.407 | 174.907 |
| St. Dev. | 3.046 | 0.532 | 2162.75 | 0.222 | 1.206 | 0.089 | 0.316 | 1.534 | 8.618 | 34.100 | 1199.770 | 1852.900 |
| Q25 | 0.467 | 0 | 6.874 | 0.042 | 0 | 0.009 | 0.076 | -0.772 | 7 | 0 | 0.759 | 1.733 |
| Median | 1.287 | 0.098 | 28.240 | 0.104 | 0.081 | 0.028 | 0.311 | -0.248 | 10 | 0 | 10.870 | 12.451 |
| Q75 | 3.386 | 0.372 | 105.280 | 0.259 | 0.319 | 0.068 | 0.665 | 0.019 | 16 | 2 | 61.514 | 54.223 |
| Panel B: Entire Compustat Sample | | | | | | | | | | | | |
| N | 174132 | 231232 | 233184 | 227049 | 233184 | 217945 | 232062 | 231783 | 275865 | 174,411 | 231653 | 240322 |
| Mean | 2.702 | 0.288 | 2810 | 0.274 | 0.070 | 0.100 | 0.200 | -0.200 | 16.800 | 5.700 | 1396.600 | 909 |
| St. Dev. | 8.107 | 0.397 | 9000.810 | 0.273 | 0.686 | 0.100 | 0.200 | 1 | 12.500 | 69.700 | 4147.600 | 5748.500 |
| Q25 | 0.296 | 0.027 | 23.483 | 0.043 | 0 | 0 | 0 | -0.100 | 8 | 0 | 12.400 | 3.900 |
| Median | 0.762 | 0.190 | 163.798 | 0.174 | 0 | 0 | 0.100 | 0 | 12 | 0 | 89.400 | 28.500 |
| Q75 | 1.786 | 0.390 | 1043 | 0.444 | 0.022 | 0.100 | 0.200 | 0.100 | 21 | 0 | 601.800 | 175.500 |

In Table 3.4 panel A, we present summary statistics that describe the average firm-level characteristics of filers in our sample. In panel B, we compare the firm characteristics of our sample to the overall sample of public firms in Compustat. Naturally, since we are studying a sample of innovative firms (i.e. firms which have at least one material IP licensing agreement), the average firm in our sample is smaller than the average firm in Compustat (mean assets of \$343M in our sample versus \$2.81B overall), more R&D intensive (mean R&D/Assets of 31.7% in our sample versus 7% overall), and younger (mean age of 13 years in our sample versus 17 years overall).

Next, we present graphical evidence that firms with greater proprietary costs of disclosure, namely firms with higher quality technology or facing more competitive product markets, are more likely to strategically redact. This is an essential condition for redaction to be a credible signal to investors, and hence justifies the positive effect of redaction on capital market and innovation outcomes as implied by our model. In Figure 3.4 panel A, we plot the fraction of redacted contracts amongst the total set of filed contracts across quintiles of product market competitiveness proxied by the Hoberg and Phillips (2016) measure of product market fluidity. This measure captures the tendency of new words in your product description to also appear in competitors' in following years, with higher values corresponding to a greater intensity of product market rivalry. The redaction rate is increasing with product market competitiveness, suggesting firms facing greater competitive threats are substantially more likely to redact. Firms in the highest quintile of product market competitiveness, who are more likely to face the threat of potential entry, redact 62% of their filings, whereas firms in the lowest quintile redact just about 10% of their filings.

In Figure 3.4 panel B, we plot the redaction rate across quintiles of total prior 5 year patent citations, for the subset of contracts whose underlying IP consists of patents. Patent citations are well-understood to be proxies for the economic value of patents (Hall, Jaffe, and Trajtenberg, 2005). Thus, contracts which license more economically valuable patents are more likely to be redacted. This effect is particularly skewed amongst the highest quintile of contracts, which have a 9% higher redaction rate than the immediately preceding quintile. The graphical evidence provides support for our argument that redaction is a signal used by firms who have more economic value

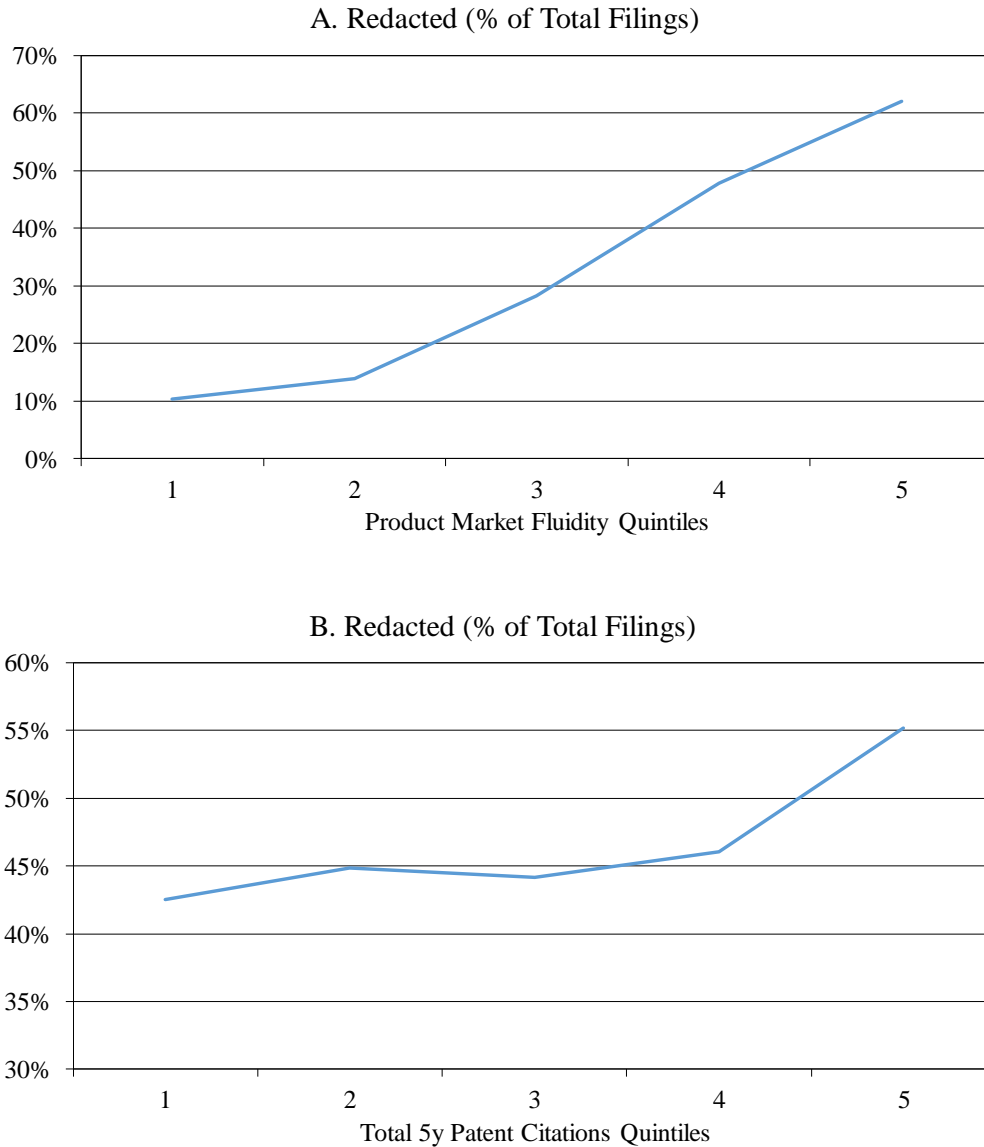


Figure 3.4: Relationship Between Redaction Rate and Proprietary Costs of Disclosure

Panel A plots the relation between quintiles of product market fluidity (the tendency of new words in your product description to also appear in competitors' in following years), as calculated by Hoberg and Phillips (2016) and the fraction of redacted contracts amongst total filed contracts in each quintile. Panel B plots the relation between quintiles of total 5 year patent citations for contracts whose underlying IP consists of patents and the fraction of redacted contracts amongst total filed contracts in each quintile.

to lose from being forced to disclose the value of their IP.

We extend this evidence with regressions explaining the probability of redaction. Table 3.5 explains the unconditional determinants of redaction, starting from baseline firm and contract level

determinants. In the first column, we focus on solely contract-level characteristics. Consistent with the argument of higher proprietary costs, contracts which are based on IP consisting of know-how, show-how, which are worldwide, and contain the right to sublicense are associated with statistically significant increases in the probability of redaction. The marginal effect of “show-how” on probability of redaction (12-13% higher probability) is almost double the effect of “know-how” (6-7% higher probability). This suggests that IP that is more difficult to codify, and thus protect (through patents, trademarks etc.), has a higher proprietary cost of disclosure and hence is associated with substantially higher probability of redaction.

In the second column, we investigate the effect of firm-level characteristics. As expected, firms with lower ROA, lower leverage, and higher intangible capital are more likely to redact. We also construct a composite variable called *SizeAge* which increases in value as a firm gets older or bigger. The exact formulation is derived from Hadlock and Pierce (2010), who argue that this index captures well the extent of financial constraints faced by a firm. We multiply this value by -1, to derive our index which we call *Smallyoung* as higher values indicate smaller or younger firms. Controlling for contract- and firm-level characteristics, however, this does not load. While these firms may be more innovative, they are also most subject to informational asymmetries and hence may receive negative capital market consequences post-redaction. We confirm that this is indeed the case in our subsequent analysis.

In the next two columns, we consider the sample of contracts filed by licensors and licensees separately. We find that licensors are substantially more likely to redact contracts with “know-how” than licensees (13.5% vs 2.2%), suggesting it is the erosion of rents for the licensor that matters most for such contracts. The effect of “show-how”, however is similar across both licensors and licensees, suggesting that the inability to protect tacit knowledge underlying the contract increases proprietary costs for both parties. The effect of the sublicense option on probability of redaction is higher for licensees than for licensors (22% vs 16%), as the option value rests largely with the licensee. Panel B indicates that firms redact when they have higher quality IP, confirming the graphical result.

Table 3.5: Benchmark Model of Determinants of Redaction

Table 5 displays estimated regression coefficients and standard errors (in brackets) from our benchmark model of contract- and firm-level determinants of redaction. The baseline regression takes the form $\mathbf{1}\{REDACT\}_{i,j,k,t} = \beta_1 X_{i,j,k,t-1} + Controls_{i,j,k,t-1} + \alpha_k + \alpha_t + \epsilon_{i,j,k,t}$. In all columns of both panels, the dependent variable is a dummy variable which takes the value of 1 if the contract was redacted and 0 otherwise. We multiply the dummy by 100 such that coefficients can be interpreted in terms of percentage probabilities. In Panel A, we consider the effects of non-payment related contract-level terms and firm-level characteristics on the likelihood of redaction. All variables are as defined in Table 1 and Appendix A.1. The model is estimated using OLS, including industry (SIC2 of the underlying technology) and year fixed effects. Standard errors (in brackets) are clustered at the technology industry (SIC2) level.

| Panel A: Contract- and Firm-Level Determinants of Redaction | | | | | |
|---|-------------|-------------|-------------|-------------|-------------|
| Dependent variable: $1\{REDACT\} \times 100$ | | | | | |
| <i>FilerLicensor</i> | 3.063 | | 2.668 | | |
| | [3.363] | | [3.514] | | |
| <i>LicensorIsIndividual</i> | -17.572*** | | -15.682*** | | -17.767*** |
| | [3.321] | | [3.198] | | [3.876] |
| <i>LicensorIsNonProfit</i> | -2.705 | | -2.521 | | -0.502 |
| | [4.3] | | [4.283] | | [4.116] |
| <i>LicensorIsUniversity</i> | -5.11** | | -1.476 | | -3.519 |
| | [2.05] | | [2.827] | | [2.505] |
| <i>IsAmendment</i> | 3.344* | | -0.18 | 4.43 | 3.126 |
| | [1.767] | | [1.571] | [3.01] | [2.07] |
| <i>IsKnowHow</i> | 7.31*** | | 6.709*** | 13.572*** | 2.205* |
| | [1.773] | | [1.738] | [4.016] | [1.271] |
| <i>IsShowHow</i> | 13.092*** | | 12.412*** | 14.146*** | 12.691*** |
| | [2.002] | | [2.281] | [1.571] | [2.11] |
| <i>Log(Duration)</i> | -0.771 | | -0.747 | -0.475 | -1.143** |
| | [0.531] | | [0.637] | [1.126] | [0.538] |
| <i>Worldwide</i> | 5.332*** | | 5.254*** | 7.133* | 4.172** |
| | [1.903] | | [1.656] | [3.917] | [1.658] |
| <i>Sublicense</i> | 20.142*** | | 19.8*** | 16.481*** | 22.314*** |
| | [3.449] | | [2.42] | [1.992] | [4.49] |
| <i>Exclusive</i> | -1.855 | | -0.607 | -1.117 | -2.793 |
| | [1.463] | | [1.513] | [2.42] | [1.773] |
| <i>ROA</i> | | -3.107*** | -1.288** | | |
| | | [1.085] | [0.62] | | |
| <i>Leverage</i> | | -3.524*** | -2.389** | | |
| | | [0.88] | [0.938] | | |
| <i>Log(KInt)</i> | | 4.675*** | 4.229*** | | |
| | | [1.113] | [1.302] | | |
| <i>SmallYoung</i> | | -0.455 | 0.54 | | |
| | | [1.129] | [1.735] | | |
| FE? | SIC2 + Year | SIC2 + Year | SIC2 + Year | SIC2 + Year | SIC2 + Year |
| Sample | Full | Full | Full | Licensor | Licensee |
| N | 5798 | 6314 | 5284 | 2283 | 3515 |
| R ² | 0.347 | 0.288 | 0.368 | 0.410 | 0.305 |

| Panel B: Patent quality and redaction | | |
|--|-------------|-------------|
| Dependent variable: $1\{REDACT\} \times 100$ | | |
| <i>5YTotCites</i> | 0.191*** | |
| | [0.046] | |
| <i>5YAvgCites</i> | | 0.061*** |
| | | [0.017] |
| FE? | SIC2 + Year | SIC2 + Year |
| Sample | Patents | Patents |
| N | 1451 | 1451 |
| $\overline{R^2}$ | 0.334 | 0.326 |

Table 3.8 explores the role of competition. Competition is one of the main proxies for proprietary costs used in the accounting literature on disclosures, and “competitive harm” is explicitly mentioned as the main justification for allowing redaction in the SEC’s guidance on requesting confidential treatment. We use three measures of competition, the *HHI*, *ProdMktFluid*, and *TSIMM*. *HHI* is computed at the Fama-French 49 industry level, based on the sample of firms appearing in Compustat. Given the known deficiencies in this measure as pointed out by Ali, Klasa, and Yeung (2009), we also use the latter two measures derived from Hoberg, Phillips, and Prabhala (2014) and Hoberg and Phillips (2016) which represent the degree of product market threat measured at the firm level, by correlating textual product descriptions from SEC filings. The first column shows that the standard concentration ratio of sales for Compustat firms in an industry does not load. However, we find that product market fluidity and textual product similarity both positively relate to HHI.

Table 3.9 presents a variety of agency measures. At the margin, we may expect firms with worse internal governance to be likelier to redact, under the hypothesis shrouding information reflects managerial agency problems. Michaely, Popadak, and Vincent (2014) define four agency measures, which we replicate. Table 3.9 indicates that across all four measures, none are *positively* related to the propensity to redact. Although one may criticize any agency measure, the direction of the estimate suggests redaction is unlikely to be due to agency problems. If anything, they are *negatively related*, significantly so in two of four specifications. In addition, to the extent competition is a substitute for governance, this suggests that agency problems are very unlikely to

Table 3.8: Effect of Competition on Likelihood of Redaction

In all columns of both panels, the dependent variable is a dummy variable which takes the value of 1 if the contract was redacted and 0 otherwise. We multiply the dummy by 100 such that coefficients can be interpreted in terms of percentage probabilities. In Panel A, we consider the effects of competition on the likelihood of redaction. *HHI* is the Herfindahl-Hirschman Index of the filing firm computed based on sales of all firms in its industry that appear in Compustat. The industry classification used in calculating *HHI* is the Fama-French 49 industries scheme. Lower values of *HHI* correspond to higher degree of competition. *ProdMktFluid* and *TSIMM* are text-based measures of product-market competition as defined in Hoberg, Phillips, and Prabhala (2014); Hoberg and Phillips (2016). Higher values of these two variables correspond to higher degree of competition. Firm controls consist of *ROA*, *Leverage*, *Log(KInt)*, and *SmallYoung*. Contract controls consist of *FilerLicensor*, *LicensorIsIndividual*, *LicensorIsNonProfit*, *LicensorIsUniversity*, *IsAmendment*, *IsKnowHow*, *IsShowHow*, *Log(Duration)*, *Worldwide*, *Sublicense*, and *Exclusive*. The model is estimated using OLS, including industry (SIC2 of the underlying technology) and year fixed effects. Standard errors (in brackets) are clustered at the technology industry (SIC2) level.

| Dependent variable: 1{ <i>REDACT</i> } × 100 | | | |
|--|-------------|-------------|-------------|
| <i>HHI</i> | 3.769 | | |
| | [31.122] | | |
| <i>ProdMktFluid</i> | | 1.661*** | |
| | | [0.305] | |
| <i>TSIMM</i> | | | 0.791*** |
| | | | [0.164] |
| <i>FilerLicensor</i> | 2.719 | 5.323* | 4.638 |
| | [3.586] | [2.786] | [3.141] |
| Firm controls? | Y | Y | Y |
| Contract controls? | Y | Y | Y |
| FE? | SIC2 + Year | SIC2 + Year | SIC2 + Year |
| N | 5230 | 3098 | 3055 |
| $\overline{R^2}$ | 0.366 | 0.394 | 0.389 |

Table 3.9: Effect of Agency Problems on Likelihood of Redaction

In this table, we consider the effects of competition on the likelihood of redaction. *Agency1*, *Agency2* and *Agency3* are measures of agency problems as defined in Michaely, Popadak, and Vincent (2014). *Agency1* is a variable that takes the value of 1 if a firm is large and has few growth opportunities, i.e., market capitalization greater than the 80th percentile and market-to-book ratio less than the 20th percentile in a given calendar year. *Agency2* refers to a dummy variable which takes the value of 1 for firms with managers that tend to overspend on SGA costs without proper economic reasons, i.e., SG&A expense greater than the 80th percentile and sales growth less than the 20th percentile in a given calendar year. *Agency3* refers to a dummy variable which takes the value of 1 if a firm had positive acquisition expenses in each of the previous two years and SDC M&A data indicates the acquisitions were not of firms in the same primary 3-digit SIC industry code as the acquirer i.e. diversifying acquisitions. $1 - Ownership$ refers to 1 minus the percentage of institutional ownership. Firm controls consist of *ROA*, *Leverage*, $\text{Log}(KInt)$, and *SmallYoung*. Contract controls consist of *FilerLicensor*, *LicensorIsIndividual*, *LicensorIsNonProfit*, *LicensorIsUniversity*, *IsAmendment*, *IsKnowHow*, *IsShowHow*, $\text{Log}(Duration)$, *Worldwide*, *Sublicense*, and *Exclusive*. The model is estimated using OLS, including industry (SIC2 of the underlying technology) and year fixed effects. Standard errors (in brackets) are clustered at the technology industry (SIC2) level.

| The Effect of Agency Problems on Decision to Redact | | | | |
|---|-------------------|----------------------|----------------------|------------------|
| Dependent variable: 1{REDACT} | | | | |
| <i>Agency</i> | 1.039 [12.109] | -6.909*** [1.231] | -9.765*** [2.089] | -0.601 [3.3] |
| <i>FilerLicensor</i> | 2.414 [3.473] | 2.735 [3.366] | 2.58 [3.586] | 2.673 [3.531] |
| Agency Var? | <i>Agency1</i> | <i>Agency2</i> | <i>Agency3</i> | $1 - Ownership$ |
| Firm Controls? | Y | Y | Y | Y |
| Contract Controls? | Y | Y | Y | Y |
| FE? | SIC2 + Year | SIC2 + Year | SIC2 + Year | SIC2 + Year |
| N | 3805 | 5269 | 5278 | 5284 |
| $\overline{R^2}$ | 0.367 | 0.370 | 0.369 | 0.367 |

be an explanation for redaction propensity.

3.5 Methodology

In this section, we present results from our empirical analysis of the effect of redaction on capital market and innovation outcomes. The results provide evidence in favor of the two key implications arising from our model and hypothesis development, first that redaction is associated with positive capital market outcomes (specifically stock liquidity, equity issuance, and institutional ownership) due to its signaling value, and second that by allowing firms to preserve IP-related competitive advantage, redaction promotes future innovation activity (specifically, patenting, intangible capital stock accumulation and R&D expenditure).

We adopt two methodologies in this section. First, we estimate annual contract-level panel regressions of the form:

$$\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t} \quad (1)$$

where Y refers to an outcome variable measured at the horizon $t+k$, where t is the contract filing year, $k=1$ year for most outcomes, and $k=2$ or 3 years for analysis of long-term consequences. $\mathbf{1}\{REDACT\}$ refers to a dummy variable for whether a given contract i , filed in year t , by firm j , with underlying technology belonging to SIC2 industry l , was redacted. \mathbf{X} is a vector of firm- and contract-level controls. We include fixed-effects for the SIC2 industry assigned to the technology underlying a given contract, α_l , and year fixed-effects, α_t .

The set of firm-level controls that we include in our regression specification consists of the logarithm of total assets ($\log(Assets)$), the logarithm of total intangible capital as defined by Peters and Taylor (2016) ($\log(K_{int})$), the ratio of cash to assets ($\frac{Cash}{Assets}$), leverage ratio ($\frac{Debt}{Assets}$), and the *SmallYoung* index value which collapses the age and size of the firm into a single index.⁶³ The set of contract-level controls includes dummy variables which take the value of 1 if the filing firm is the licensor (*FilerLicensor*), if the licensor is an individual (*LicensorIsIndividual*), a non-profit (*LicensorIsNonProfit*), a university (*LicensorIsUniversity*), if the given contract is an amendment of a previous agreement (*IsAmendment*), if the underlying IP consists of know-how⁶⁴ (*IsKnowHow*), and show-how⁶⁵ (*IsShowHow*). We also include the logarithm of the duration of the contract ($\log(Duration)$), a dummy for whether the licensing rights extend worldwide (*Worldwide*), whether the agreement confers the licensee the right to sublicense the IP (*Sublicense*), and if the IP is being licensed exclusively to that particular licensee (*Exclusive*). We include these variables in order to capture common economic determinants of licensing con-

⁶³The *SmallYoung* index is calculated as minus one times the *SizeAge* index as computed by Hadlock and Pierce (2010). The *SmallYoung* index is calculated as $-1 \times SizeAge = -1 \times [(-0.744 \times Size) + (0.042 \times Size^2) - (0.075 \times Age) + (0.001 \times Age^2)]$, where *Size* is the log of inflation-adjusted book assets, and *Age* is the number of years the firm is listed with a non-missing stock price on Compustat. As in Hadlock and Pierce (2010), when we construct the index, *Size* is winsorized at $\log(4.5 \times 10^9)$, and *Age* is winsorized at 37 years.

⁶⁴Know-how encompasses unpatentable proprietary information including blueprints, drawings, data etc.

⁶⁵Show-how consists of proprietary assistance rendered by the licensor to the licensee.

tract terms (as noted in Varner (2011) and Hegde (2014)).

Our coefficient of interest in equation (1) is β_1 , and it can be interpreted as a “difference-in-difference”. Specifically, when $k = 1$, the coefficient β_1 captures the difference in the change in variable Y from time t (the filing date) to $t + 1$, denoted as $\Delta Y_{t+1} = Y_{t+1} - Y_t$, between firms which redact and firms which do not. A positive estimate of β_1 indicates that ΔY_{t+1} for redacting firms is on average greater than ΔY_{t+1} for non-redacting firms. We control for the current level of the dependent variable Y_t to mitigate any scale effects not already captured by controlling for total assets and intangible capital.

Second, we employ matching techniques in order to improve comparability between the sample of firms which redact and those which do not. We match redacted and un-redacted contracts on the basis of a propensity score which is a function of the *SmallYoung* index (size and age being important unconditional determinants of a firm’s likelihood to file patents) and the level of intangible capital ($\log(K_{int})$). In addition, we require an exact match on the 1-digit SIC industry of the licensed technology, and we require that the treated and untreated contracts fall within a 3 year block of each others’ filing dates. Finally, we require an exact match on whether the underlying IP consists of patents or not, as firms which patent ex-ante are much more likely to continue patenting. We evaluate the improvement in comparability of the samples generated by the matching procedure by considering the percent improvement in balance for each of the matching variables, defined as $100 \times \frac{(|a| - |b|)}{|a|}$, where a is the balance statistic before and b is the balance statistic after matching. The balance statistics considered include means of each matching variable, as well as mean, median and maximum value of the empirical distribution of each matching variable across the treatment and control samples. In Table 3.10, we show the mean improvement in balance for each of the variables on which we construct the propensity score, and the score itself, as well as improvement in the mean, median, and maximum of the empirical distribution of each matching variable across the treatment and control samples. The score itself (labeled as Distance), improves by 99% on average, the balance of the *SmallYoung* index improves by 70% on average, and the balance of intangible capital improves by 94% on average. As noted by Imai, King, and Stuart

Table 3.10: Percentage Improvement in Balance as a Result of Matching

In this table, we evaluate the improvement in comparability of the samples generated by matching by considering the percent improvement in balance for each of the matching variables, defined as $100 \times \frac{(|a| - |b|)}{|a|}$, where a is the balance statistic before and b is the balance statistic after matching for a given variable. This table shows the mean improvement in balance for each of the variables on which we construct the propensity score, and the score itself as well as quantiles of the empirical distribution of the improvement in balance across all observations in the matched sample. *Distance* is the propensity score. All other variables are as defined in the text.

| | Mean Diff. | eQQ Med | eQQ Mean | eQQ Max |
|-------------------|------------|---------|----------|---------|
| <i>Distance</i> | 99.988 | 86.656 | 87.084 | 42.903 |
| <i>SmallYoung</i> | 70.312 | 30.984 | 48.566 | 16.833 |
| <i>log(KInt)</i> | 94.277 | 93.664 | 88.578 | 59.638 |

(2008), this is a superior way of comparing balance of covariates pre- and post-matching.⁶⁶ Finally, in the matching analysis, we cluster the standard errors at the level of the matches, as recommended by Abadie and Spiess (2016b).⁶⁷

We interpret the results of the OLS and matching analysis jointly in order to verify the two key implications of our model. The first implication is that redaction is associated with positive capital market outcomes. The capital market outcomes we consider include stock liquidity (measured by the Amihud (2002) illiquidity measure, bid-ask spread, and turnover, as well as percentage price impact in our robustness analysis), the likelihood of equity issuance (SEO), and institutional ownership (measured by percentage institutional ownership, percentage owned by top 5 institutional owners, and concentration of institutional ownership). The second implication is that redaction is associated with positive innovation outcomes. The innovation outcomes we consider include patenting (measured by total citation-weighted patents and total stock market value associated with patents) granted to a given firm in a given year, research and development spending, and intangible capital accumulation (as defined by Peters and Taylor (2016)).

After establishing the main result, we then turn to the cross-section. We would expect adverse

⁶⁶The usual method of computing t-statistics on the difference in means of the variables of interest is highly misleading for four reasons described in their paper.

⁶⁷The broad inferences are robust to alternative matching schemes, but we choose a somewhat coarse industry definition and 3 year time intervals to reduce oversampling a subset of contracts in the control group.

selection concerns to limit the ability of firms to signal their technology's type. We therefore explore cross-sectional heterogeneity in effect of redaction on stock liquidity, and test whether interaction with variables associated with high information asymmetry and low reputation firms generates a negative marginal relationship between redaction and liquidity. First, we test whether the liquidity effects are partially undone for small and young firms, for whom information asymmetries are likely to be more acute. Second, we test whether firms which have a higher rate of past redaction also face a negative marginal effect of redaction on liquidity. That is, we check if there is a degree of concavity in the extent to which firms can credibly signal high quality technology through redaction. Finally, we test whether firms with greater equity dependence, who due to repeated interactions with investors are less likely to be affected by adverse selection in capital markets, experience a positive marginal effect of redaction on stock liquidity.

3.6 Consequences of Strategic Redaction

3.6.1 Redaction and Capital Market Outcomes

In Table 3.11, we consider the effect of redaction on future stock liquidity measured three ways. For each contract i filed by firm j at time t , we compute daily values of Amihud's illiquidity measure (Amihud, 2002) for each day in the 1 year after time t . We do this by computing the ratio of the daily absolute return to the (dollar) trading volume on that day, for firm j 's stock. We average this value across all days in the 12 months after filing time t , to obtain the variable *Illiq*. We also compute daily bid-ask spread for firm j 's stock, which is the closing ask minus the closing bid scaled by the end-of-day price, and average this across all days in the 12 months after filing time t to obtain the variable *Spread*. Finally, we compute daily turnover for firm j 's stock, which is the end-of-day price multiplied by the volume of shares traded. We average this across all days in the 12 months after filing time t to obtain the variable *Turnover*. With these three outcome variables, we estimate the panel regression in equation (1). Results are reported in Table 3.11.

The first three columns contain OLS coefficient estimates for specification (1) with firm- and

Table 3.11: Effect of Redaction on Future Stock Liquidity

Table 6 displays estimated regression coefficients and standard errors (in brackets) corresponding to the effects of redaction on future liquidity. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. For each contract i filed by firm j at time t , we compute daily values of Amihud's illiquidity measure (Amihud, 2002) for each day in the 1 year after time t . We do this by computing the ratio of the daily absolute return to the (dollar) trading volume on that day, for firm j 's stock. We average this value across all days in the 12 months after filing time t , to obtain the variable *Illiq*. We also compute daily bid-ask spread for firm j 's stock, which is the closing ask minus the closing bid scaled by the end-of-day price, and average this across all days in the 12 months after filing time t to obtain the variable *Spread*. Finally, we compute daily turnover for firm j 's stock, which is the end-of-day price multiplied by the volume of shares traded. We average this across all days in the 12 months after filing time t to obtain the variable *Turnover*. The first three columns contain OLS coefficient estimates from specifications which include with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level. The second three columns contain estimates from a matched sample. Matching is done with replacement based on a propensity score which is a function of *SmallYoung* and $\log(KInt)$, plus exact matching on SIC1 of the technology, filing dates within a 3 year window of each other, and underlying IP consisting of patents. Standard errors in the matching analysis are clustered at the level of the matches as recommended by Abadie and Spiess (2016b).

| | $\Delta Illiq_{t+1}$ | $\Delta Spread_{t+1}$ | $\Delta Turnover_{t+1}$ | $\Delta Illiq_{t+1}$ | $\Delta Spread_{t+1}$ | $\Delta Turnover_{t+1}$ |
|--------------------------|----------------------|-----------------------|-------------------------|----------------------|-----------------------|-------------------------|
| $\mathbf{1}\{REDACT\}$ | -0.013* | -0.042*** | 0.072*** | -0.011** | -0.044*** | 0.060*** |
| | [0.007] | [0.009] | [0.021] | [0.005] | [0.011] | [0.017] |
| $Level_t$ | -0.179*** | -0.044*** | -0.024*** | -0.124*** | -0.054*** | -0.021*** |
| | [0.015] | [0.004] | [0.004] | [0.020] | [0.005] | [0.002] |
| (Intercept) | | | | 0.034*** | 0.072*** | 0.120*** |
| | | | | [0.004] | [0.009] | [0.016] |
| Analysis | OLS | OLS | OLS | Matching | Matching | Matching |
| SIC2 + Year FE | Y | Y | Y | N | N | N |
| Contract + Firm Controls | Y | Y | Y | N | N | N |
| N | 3324 | 3303 | 3324 | 2632 | 2621 | 2632 |
| $\overline{R^2}$ | 0.235 | 0.264 | 0.200 | 0.062 | 0.087 | 0.072 |

contract-level controls, as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology industry level. The second three columns contain estimates from a matched sample. The coefficient of interest is the coefficient on $\mathbf{1}\{REDACT\}$, which measures the difference in the growth of the dependent variable in the 1 year post-filing, between firms which file redacted and un-redacted contracts. For $\Delta Illiq$ and $\Delta Spread$, across all specifications, the coefficients on $\mathbf{1}\{REDACT\}$ are negative and significant, indicating redaction is associated with improved liquidity post-filing, whereas the coefficient for $\Delta Turnover$ is positive. These magnitudes are economically significant. For instance, the mean value of *Illiq* in our sample is 0.1776, and thus the marginal effect of redaction of -0.013 from the first column indicates that firms which file redacted contracts are associated with a post-filing improvement in *Illiq* of 7% of the mean, relative to firms which file un-redacted contracts. In the Appendix (Table C.4.1), we also report results in event-time that indicate liquidity (measured by price impact and effective spread) improves in the 2 weeks immediately post-filing, and in the 15 weeks post-filing, for redacting relative to non-redacting firms. Thus, at the short-, mid-, and long-term horizons, liquidity improves post-filing for firms filing redacted contracts relative to those filing un-redacted ones. This suggests that contrary to the findings of Verrecchia and Weber (2006), investors do not penalize firms which strategically redact disclosures related to IP licensing agreements through reduced liquidity.⁶⁸ These results provide support for the first key implication of our model, that redaction serves as a credible, investor-recognized signal of technology type, and is associated with a positive capital market response. All results from our main analyses, and indeed the vast majority of results from our cross-sectional analyses, obtain even after including technology industry-by-year fixed effects. We present a selection of these in the Appendix Table C.4.2.

⁶⁸This can be partially attributed to the fact that Verrecchia and Weber (2006)'s sample differs from ours in two key ways. First, they focus on small firms (between \$50 and 100 million market capitalization). In contrast, the average firm in our sample has a market capitalization of \$990 million, and the largest firm in our sample has a market capitalization of \$279 billion. Second, they consider all varieties of redacted material agreements, including debt, equity, employment, customer-supplier, and IP licensing related agreements. Our sample focuses exclusively on IP licensing agreements, and thus includes relatively innovative firms, for whom the benefits of non-disclosure are likely to be more pronounced. Consistent with their findings, however, we show that small and young firms that redact in our sample, too suffer a marginally negative effect on their stock liquidity post-filing as compared to firms which do not redact.

Table 3.12: Effect of Redaction on Future Equity Issuance and Institutional Ownership

This table displays estimated regression coefficients and standard errors (in brackets) corresponding to the effects of redaction on equity issuance and institutional ownership. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. $\mathbf{1}\{Issuance\}_t$ is a dummy variable that takes the value of 1 if a given firm conducts a secondary equity offering (SEO), in which capital is raised through an equity issuance, in a given year t . $InstOwn$ refers to the percentage of outstanding shares owned by institutions (entities that file 13/Fs), $Top5Share$ refers to the percentage of shares owned by the top 5 institutional owners, out of all shares owned by institutional investors, and HHI refers to the Herfindahl-Hirschman index of institutional ownership (with lower values implying lower concentration). The first three columns contain OLS coefficient estimates from specifications with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level. The second three columns contain estimates from a matched sample. Matching is done with replacement based on a propensity score which is a function of $SmallYoung$ and $\log(KInt)$, plus exact matching on SIC1 of the technology, filing dates within a 3 year window of each other, and underlying IP consisting of patents. Standard errors in the matching analysis are clustered at the level of the matches as recommended by Abadie and Spiess (2016b).

| Panel A: Effect of Redaction on Equity Issuance | | | | | | |
|---|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| | $\mathbf{1}\{Issuance\}_{t+1}$ | $\mathbf{1}\{Issuance\}_{t+2}$ | $\mathbf{1}\{Issuance\}_{t+3}$ | $\mathbf{1}\{Issuance\}_{t+1}$ | $\mathbf{1}\{Issuance\}_{t+2}$ | $\mathbf{1}\{Issuance\}_{t+3}$ |
| $\mathbf{1}\{REDACT\}$ | 0.022*** [0.007] | 0.032** [0.016] | 0.029 [0.020] | 0.057*** [0.011] | 0.104*** [0.013] | 0.127*** [0.011] |
| Analysis | OLS | OLS | OLS | Matching | Matching | Matching |
| SIC2 + Year FE | Y | Y | Y | N | N | N |
| Contract + Firm Controls | Y | Y | Y | N | N | N |
| N | 4943 | 4943 | 4943 | 3664 | 3664 | 3664 |
| $\overline{R^2}$ | 0.083 | 0.111 | 0.127 | 0.020 | 0.020 | 0.020 |

| Panel B: Effect of Redaction on Institutional Ownership | | | | | | |
|---|------------------------|--------------------------|----------------------|------------------------|--------------------------|----------------------|
| | $\Delta InstOwn_{t+1}$ | $\Delta Top5Share_{t+1}$ | ΔHHI_{t+1} | $\Delta InstOwn_{t+1}$ | $\Delta Top5Share_{t+1}$ | ΔHHI_{t+1} |
| $\mathbf{1}\{REDACT\}$ | 0.031*** [0.006] | 21.809*** [3.147] | -0.037*** [0.009] | 0.043*** [0.006] | 25.002*** [3.203] | -0.017** [0.007] |
| $Level_t$ | -0.081*** [0.005] | -0.416*** [0.021] | -0.839*** [0.031] | -0.216*** [0.009] | -0.360*** [0.012] | -0.544*** [0.026] |
| (Intercept) | | | | 0.098*** [0.005] | 76.258*** [3.054] | 0.144*** [0.007] |
| Analysis | OLS | OLS | OLS | Matching | Matching | Matching |
| SIC2 + Year FE | Y | Y | Y | N | N | N |
| Contract + Firm Controls | Y | Y | Y | N | N | N |
| N | 4674 | 4674 | 4675 | 3390 | 3390 | 3390 |
| $\overline{R^2}$ | 0.192 | 0.260 | 0.475 | 0.140 | 0.248 | 0.252 |

We provide further support for this implication by considering the effect of redaction on post-filing likelihood of equity issuance and institutional ownership. In Table 3.12 panel A, we consider the effect of a firm redacting its filing on the likelihood that it conducts a seasoned equity offering (SEO) in the subsequent 1, 2, and 3 years. The first three columns contain OLS coefficient estimates, and the second three columns contain matching results. The coefficients on $\mathbf{1}\{REDACT\}$ indicate that redacting firms have a 2.2% higher likelihood of an SEO in the next 1 year, 3.2% higher in 2 years, and 2.9% higher in 3 years, relative to firms which do not redact their filings. The matching results provide even greater economic magnitudes, to the tune of 5.7% to 12.7% greater likelihood of SEOs post-filing for redacting firms relative to those which do not redact their filings. Thus, not only do investors reward redacting firms with improved liquidity, these firms are also more likely to issue equity in the future. This further supports the implication that redaction serves as a credible signal of the value of the underlying technology to investors.

In Table 3.12 panel B, we measure the effect of redaction on institutional ownership. We use three proxies for institutional ownership. *InstOwn* refers to the percentage of outstanding shares owned by institutions (entities that file 13/Fs), *Top5Share* refers to the percentage of shares owned by the top 5 institutional owners, out of all shares owned by institutional investors, and *HHI* refers to the Herfindahl-Hirschman index of institutional ownership (with lower values implying lower concentration). The results from the first and fourth columns indicate that overall institutional ownership increases subsequent to filing for redacting firms relative to non-redacting firms. The positive and significant effect of redaction on $\Delta Top5Share$ indicates that post-filing, the major existing institutional owners expand their holdings in redacting firms relative to those which do not redact. Finally, the negative and significant effect on ΔHHI indicates that concentration of institutional ownership declines post-filing for firms which redact relative to those which do not. This implies that both existing owners and new owners recognize the need for innovative firms to redact their disclosures strategically, to avoid competitive harm. The decrease in concentration suggests that it is *not only* existing owners, who may have access to private information, who expand their holdings in redacting firms post-filing. Thus, rather than deterring them, redaction

signals the presence of a valuable technology and leads to an expansion in both existing and new institutional ownership.

Taken together, the results from this section provide counter-evidence to the logic that redaction is perceived negatively by investors. While redaction mechanically deprives investors of information, the ability of firms to signal the value of their technology by redacting is recognized and rewarded by investors through improved stock liquidity, greater equity issuance, and institutional ownership.

3.6.2 Redaction and Future Innovation

The next set of consequences we consider are related to future innovation activity. In Table 3.13, panel A we display regressions in which the dependent variables are various measures of patenting at a 1 year-ahead horizon with respect to the filing date.

The first three columns contain full-sample OLS estimates, and the last three columns contain matched sample estimates. We use two measures of patenting, borrowed from Kogan et al. (2016). *TCW* refers to the forward citation weighted count of all patents granted to a given firm in a given year, with adjustments for citation truncation lags. *TSM* refers to the total dollar value of patents granted to a given firm in a given year, estimated from the stock market reaction to the patent grants. Our outcome variable is therefore the logarithm of the innovation output, which can be interpreted as an intensive margin test. However, un-tabulated extensive margin tests also produce directionally similar results. Across all specifications, $\mathbf{1}\{REDACT\}$ loads significantly and positively, suggesting firms which strategically redact a given contract experience increased patenting in the 1 year post-filing, measured both in terms of a citation-weighted count, and in terms of stock market value generated, relative to non-redacting firms. The effect is obtained in both the matched sample and through OLS estimates, controlling for contract- and firm-level characteristics, including the firm's prior patenting and intangible capital stock. Furthermore, when we allow for asymmetric effects between the licensor and the licensee, the licensor seems to have a stronger marginal effect in terms of even higher future patenting, as seen in the third and sixth

Table 3.13: Effect of Redaction on Future Patenting

Panel A displays estimated regression coefficients and standard errors (in brackets) corresponding to the effects of redaction on future patenting at the 1 year ahead horizon. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. We use two measures of patenting, borrowed from Kogan et al. (2016). *TCW* refers to the forward citation weighted count of all patents granted to a given firm in a given year, with adjustments for citation truncation lags. *TSM* refers to the total dollar value of patents granted to a given firm in a given year, estimated from the stock market reaction to the patent grants. The first three columns contain OLS coefficient estimates from specifications with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level. The second three columns contain estimates from a matched sample. Matching is done with replacement based on a propensity score which is a function of *SmallYoung* and $\log(KInt)$, plus exact matching on SIC1 of the technology, filing dates within a 3 year window of each other, and underlying IP consisting of patents. Standard errors in the matching analysis are clustered at the level of the matches as recommended by Abadie and Spiess (2016b).

| Panel A: Effect of Redaction on Future Innovation (1 year ahead) | | | | | | |
|--|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| | $\Delta \log(TCW)_{t+1}$ | $\Delta \log(TSM)_{t+1}$ | $\Delta \log(TSM)_{t+1}$ | $\Delta \log(TCW)_{t+1}$ | $\Delta \log(TSM)_{t+1}$ | $\Delta \log(TSM)_{t+1}$ |
| $\mathbf{1}\{REDACT\}$ | 0.413*** | 0.404*** | 0.263*** | 0.344*** | 0.414*** | 0.182*** |
| | [0.070] | [0.090] | [0.064] | [0.044] | [0.049] | [0.060] |
| $\mathbf{1}\{REDACT\} \times FilerLicensor$ | | | 0.331*** | | | 0.523*** |
| | | | [0.095] | | | [0.107] |
| <i>FilerLicensor</i> | | | 0.103*** | | | 0.033 |
| | | | [0.040] | | | [0.071] |
| <i>FilerLicensor</i> \times $\log(TotalPriorPatents)$ | | | 0.040 | | | -0.021 |
| | | | [0.041] | | | [0.036] |
| $\log(TotalPriorPatents)$ | 0.356*** | 0.414*** | 0.384*** | 0.465*** | 0.567*** | 0.564*** |
| | [0.020] | [0.029] | [0.033] | [0.015] | [0.018] | [0.023] |
| (Intercept) | | | | 0.381*** | 0.131*** | 0.134*** |
| | | | | [0.038] | [0.033] | [0.041] |
| Analysis | OLS | OLS | OLS | Matching | Matching | Matching |
| SIC2 + Year FE | Y | Y | Y | N | N | N |
| Contract + Firm Controls | Y | Y | Y | N | N | N |
| N | 3316 | 3316 | 3316 | 2438 | 2438 | 2438 |
| $\overline{R^2}$ | 0.457 | 0.520 | 0.530 | 0.333 | 0.389 | 0.404 |

Panel B displays estimated regression coefficients and standard errors (in brackets) corresponding to the effects of redaction on future patenting at the 2 to 3 year ahead horizon. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. TSM refers to the total dollar value of patents granted to a given firm in a given year, estimated from the stock market reaction to the patent grants. The first two columns contain OLS coefficient estimates from specifications with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level. The second two columns contain estimates from a matched sample. Matching is done with replacement based on a propensity score which is a function of $SmallYoung$ and $\log(KInt)$, plus exact matching on SIC1 of the technology, filing dates within a 3 year window of each other, and underlying IP consisting of patents. Standard errors in the matching analysis are clustered at the level of the matches as recommended by Abadie and Spiess (2016b).

| Panel B: Effect of Redaction on Future Innovation (2-3 years ahead) | | | | |
|---|--------------------------|--------------------------|--------------------------|--------------------------|
| | $\Delta \log(TSM)_{t+2}$ | $\Delta \log(TSM)_{t+3}$ | $\Delta \log(TSM)_{t+2}$ | $\Delta \log(TSM)_{t+3}$ |
| $\mathbf{1}\{REDACT\}$ | 0.308** [0.120] | 0.404*** [0.109] | 0.383*** [0.055] | 0.501*** [0.059] |
| $\log(TotalPriorPatents)$ | 0.391*** [0.027] | 0.354*** [0.033] | 0.578*** [0.022] | 0.563*** [0.024] |
| (Intercept) | | | 0.115*** [0.037] | 0.107*** [0.040] |
| Analysis | OLS | OLS | Matching | Matching |
| SIC2 + Year FE | Y | Y | N | N |
| Contract + Firm Controls | Y | Y | N | N |
| N | 2979 | 2669 | 2181 | 1972 |
| $\overline{R^2}$ | 0.493 | 0.473 | 0.378 | 0.350 |

columns. This reflects the role of the licensor as the entity generating the IP - a licensee may build follow-on IP, or simply utilize this innovation in a downstream application.

The magnitudes of the estimated effects are striking in their economic significance. The mean value of *TCW* is 0.8204, and the estimate from the first column of 0.4126 implies that redacting firms have an increase in citation-weighted patenting representing around 50% of the mean, relative to firms which do not redact. Similarly, the estimated effect of redaction from the second column is around 53% of the mean value of *TSM*. The magnitudes from the matched sample are lower, but still significant, representing 40-50% of the mean of the respective variables.

In panel B of Table 3.13, we show regressions estimates that measure the effects of strategic redaction on patenting at a longer horizon of 2 and 3 years after the filing date. The estimates of the coefficients on $\mathbf{1}\{REDACT\}$ are of similar statistical and economic significance as those in panel A of Table 3.13. Redacting firms thus seem to follow a sustained, higher innovation path post-filing when compared with non-redacting firms.

To the extent that patenting is an incomplete measure of firm-level innovation activity, we also consider the effect of strategic redaction on the growth of various forms of intangible capital as defined in Peters and Taylor (2016). Although R&D spending is not *necessarily* innovation output, the *growth* of spending itself represents innovation effort, and innovation output is likely increasing in innovation effort. We consider the following four measures. First, the firm's expenditure on research and development divided by total assets ($\frac{R\&D}{Assets}$), second, the level of intangible capital (*KInt*). This variable is provided through WRDS, and is defined in Peters and Taylor (2016) as the sum of on and off balance sheet intangible assets, including reported intangible assets that appear on the balance sheet, and accumulation of R&D expenses, and a fraction of SG&A expenses using the perpetual inventory method. They also calculate disaggregated components of intangible capital stock including knowledge capital, or *KKnow*, which is calculated as the accumulation of R&D expenses using the perpetual inventory method, and off-balance sheet capital, or *KOffBS*, which includes the accumulation of R&D and a portion of SG&A expenses through the perpetual inventory method. The latter measure incorporates the portion of a firm's intangible assets that do

not appear on the balance sheet since these are expenses listed on the income statement. Results shown in Table 9 indicate a positive and significant coefficient of $\mathbf{1}\{REDACT\}$ on all the four measures, suggesting that post-filing, redacting firms spend more on R&D and accumulate more intangible capital than non-redacting firms.

On the whole, strategic redaction of an IP licensing agreement predicts significantly greater future innovation post-filing based on both patent and non-patent measures. Taken together, these results provide strong evidence in favor of the second key implication of our model, that by allowing high technology-quality firms to redact enables them to preserve comparative advantage and prevent competitive entry into their technology space, thereby promoting future innovation activity.

3.6.3 Cross-sectional Heterogeneity

In this section, we analyze the cross-sectional heterogeneity in the effect of redaction on capital market outcomes. We propose three tests to capture the intuition of the hypothesis that the positive capital market effect is muted for firms with a less credible reputation amongst investors. In order to perform these tests, we modify the initial specification in equation (1) to allow for cross-sectional heterogeneity in the effect of redaction:

$$\begin{aligned} \Delta Y_{i,j,l,t+k} = & \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \beta_3 Z_{i,j,l,t} \\ & + \beta_4 \mathbf{1}\{REDACT\}_{i,j,l,t} \times Z_{i,j,l,t} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t} \end{aligned} \quad (2)$$

where all variables are as previously defined, except that now we include an additional term which interacts the main effect, $\mathbf{1}\{REDACT\}$ with a cross-sectional variable Z .

First, we study whether the effect of redaction on capital market outcomes varies with the size and age of the filing firm. These attributes are collapsed into a single index, which we denote *SmallYoung* (defined previously), which decreases in value as a firm grows in size or age. The

Table 3.14: Effect of Redaction on Future Intangible Capital Accumulation and R&D Expense

This table displays estimated regression coefficients and standard errors (in brackets) corresponding to the effects of redaction on future intangible capital accumulation. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. We consider the following four measures. First, the firm's expenditure on research and development divided by total assets ($\frac{R\&D}{Assets}$), second, the level of intangible capital ($KInt$). This variable is provided through WRDS, and is defined in Peters and Taylor (2016) as the sum of on and off balance sheet intangible assets, including reported intangible assets that appear on the balance sheet, and accumulation of R&D expenses, and a fraction of SG&A expenses using the perpetual inventory method. They also calculate disaggregated components of intangible capital stock including knowledge capital, or $KKnow$, which is calculated as the accumulation of R&D expenses using the perpetual inventory method, and off-balance sheet capital, or $KOffBS$, which includes the accumulation of R&D and a portion of SG&A expenses through the perpetual inventory method. The first four columns contain OLS coefficient estimates from specifications with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level. The second four columns contain estimates from a matched sample. Matching is done with replacement based on a propensity score which is a function of $SmallYoung$ and $\log(KInt)$, plus exact matching on SIC1 of the technology, filing dates within a 3 year window of each other, and underlying IP consisting of patents. Standard errors in the matching analysis are clustered at the level of the matches as recommended by Abadie and Spiess (2016b).

| Effect of Redaction on Intangible Capital Accumulation and R&D Expense | | | | | | | | |
|--|------------------------------------|---------------------|----------------------|-----------------------|------------------------------------|---------------------|----------------------|-----------------------|
| | $\Delta \frac{R\&D}{Assets}_{t+1}$ | $\Delta KInt_{t+1}$ | $\Delta KKnow_{t+1}$ | $\Delta KOffBS_{t+1}$ | $\Delta \frac{R\&D}{Assets}_{t+1}$ | $\Delta KInt_{t+1}$ | $\Delta KKnow_{t+1}$ | $\Delta KOffBS_{t+1}$ |
| $\mathbf{1}\{REDACT\}$ | 0.137*** | 0.166*** | -0.021 | 0.112*** | 0.032*** | 0.201*** | 0.244*** | 0.206*** |
| | [0.023] | [0.018] | [0.073] | [0.024] | [0.012] | [0.019] | [0.020] | [0.016] |
| (Intercept) | | | | | -0.022** | 0.376*** | 0.315*** | 0.360*** |
| | | | | | [0.009] | [0.015] | [0.016] | [0.012] |
| Analysis | OLS | OLS | OLS | OLS | Matching | Matching | Matching | Matching |
| SIC2 + Year FE | Y | Y | Y | Y | N | N | N | N |
| Contract + Firm Controls | Y | Y | Y | Y | N | N | N | N |
| N | 3005 | 3952 | 3952 | 3952 | 2401 | 2704 | 2704 | 2704 |
| $\overline{R^2}$ | 0.017 | 0.114 | 0.058 | 0.131 | 0.002 | 0.035 | 0.046 | 0.049 |

Table 3.15: Effect of Redaction for Small and Young Firms

This table displays estimated regression coefficients and standard errors (in brackets) corresponding to the cross-sectional effects of firm size and age. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \beta_3 Z_{i,j,l,t} + \beta_4 \mathbf{1}\{REDACT\}_{i,j,l,t} \times Z_{i,j,l,t} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. All variables are as previously defined. The regressions are estimated by OLS with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level.

| | $\Delta \log(TSM)_{t+1}$ | $\Delta \log(TSM)_{t+2}$ | $\Delta Illiq_{t+1}$ | $\Delta Spread_{t+1}$ |
|--|--------------------------|--------------------------|----------------------|-----------------------|
| $\mathbf{1}\{REDACT\}$ | 0.326*** [0.085] | 0.229** [0.115] | -0.019*** [0.007] | -0.051*** [0.011] |
| $\mathbf{1}\{REDACT\} \times SmallYoung$ | 0.323*** [0.077] | 0.244*** [0.091] | 0.032*** [0.007] | 0.047*** [0.018] |
| <i>SmallYoung</i> | -1.079*** [0.249] | -1.229*** [0.321] | 0.001 [0.027] | 0.032 [0.020] |
| <i>Level_t</i> | | | -0.181*** [0.004] | -0.045*** [0.004] |
| $\log(TotalPriorPatents)$ | 0.418*** [0.026] | 0.398*** [0.022] | | |
| Analysis | OLS | OLS | OLS | OLS |
| SIC2 + Year FE | Y | Y | Y | Y |
| Contract + Firm Controls | Y | Y | Y | Y |
| N | 3316 | 2979 | 3324 | 3303 |
| $\overline{R^2}$ | 0.556 | 0.537 | 0.238 | 0.267 |

results are contained in Table 10. The estimated coefficients on $\mathbf{1}\{REDACT\} \times SmallYoung$ indicate that while smaller and younger firms that redact do innovate more post-filing as measured by stock market value generated by patent grants in the 1 year post-filing (*TSM*), their liquidity, measured by Amihud illiquidity (*Illiq*) and bid-ask spread (*Spread*), worsens post-filing relative to non-redacting firms. The smaller or younger the firm, as measured by the *SmallYoung* index value, the more negative the relationship between redaction and liquidity. This is consistent with the findings of Verrecchia and Weber (2006) and Boone, Floros, and Johnson (2016), who both document that redaction is negatively associated with liquidity in samples of firms that are skewed towards smaller and/or younger firms. Specifically, the results found by Verrecchia and Weber (2006) are nested within ours, as the negative effects of redaction on liquidity that they document are driven by the fact that their sample consists of only small firms (market capitalization between

Table 3.16: Effect of Redaction Conditional on Past Redaction

This table displays estimated regression coefficients and standard errors (in brackets) corresponding to the cross-sectional effects of firms' past redaction tendency. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \beta_3 Z_{i,j,l,t} + \beta_4 \mathbf{1}\{REDACT\}_{i,j,l,t} \times Z_{i,j,l,t} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. We measure prior redaction tendency in four ways. First, $\mathbf{1}\{PriorRedacter\}$ is a dummy variable which takes the value 1 if a firm has previously filed a redacted agreement in our sample, prior to time t . Second, $\log(DaysSinceFirstRedact)$ captures the number of days since the first redacted agreement filed by the firm in question was filed. To the extent that we do not capture all redacted filings by a given firm, we assume that this measure captures prior redaction history assuming a uniform within-firm rate of redaction over time. Third, the variable $\log(\#PriorRedact)$ captures the number of prior redacted agreements filed by the given firm, prior to the filing of the current agreement. Fourth, the variable $RedactRate$ is simply the fraction of the number of filings by the given firm which were redacted divided by the total number of filings (both redacted and un-redacted) submitted by the given firm, prior to the filing of the current agreement. All other variables are as previously defined. The regressions are estimated by OLS with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level.

| | $\Delta Spread_{t+1}$ | $\Delta Spread_{t+1}$ | $\Delta Spread_{t+1}$ | $\Delta Spread_{t+1}$ |
|---|-----------------------|-----------------------|-----------------------|-----------------------|
| $\mathbf{1}\{REDACT\}$ | -0.060*** [0.011] | -0.048*** [0.013] | -0.061*** [0.008] | -0.070*** [0.008] |
| $\mathbf{1}\{REDACT\} \times \mathbf{1}\{PriorRedacter\}$ | 0.053* [0.027] | | | |
| $\mathbf{1}\{REDACT\} \times \log(DaysSinceFirstRedact)$ | | | | 0.010*** [0.003] |
| $\mathbf{1}\{REDACT\} \times \log(\#PriorRedact)$ | | | 0.025** [0.012] | |
| $\mathbf{1}\{REDACT\} \times RedactRate$ | | 0.058*** [0.020] | | |
| $\mathbf{1}\{PriorRedacter\}$ | -0.037* [0.020] | | | |
| $\log(DaysSinceFirstRedact)$ | | | | -0.004 [0.003] |
| $\log(\#PriorRedact)$ | | | -0.015 [0.010] | |
| $RedactRate$ | | -0.066*** [0.012] | | |
| $Level_t$ | -0.044*** [0.004] | -0.044*** [0.004] | -0.045*** [0.004] | -0.045*** [0.004] |
| Analysis | OLS | OLS | OLS | OLS |
| SIC2 + Year FE | Y | Y | Y | Y |
| Contract + Firm Controls | Y | Y | Y | Y |
| N | 3303 | 3303 | 3303 | 3303 |
| $\overline{R^2}$ | 0.265 | 0.265 | 0.265 | 0.266 |

\$50 to \$100 million). We find that while smaller and younger innovative firms that redact are no less innovative, they are met with a negative capital market response in terms of liquidity. This suggests that reputation and ex-ante perceived information asymmetry moderates the degree to which investors perceive redaction as a credible signal.

Second, we analyze whether firms lose their ability to credibly signal high value technology through redaction if they have frequently redacted in the past. We measure prior redaction tendency in four ways. First, $\mathbf{1}\{PriorRedacter\}$ is a dummy variable which takes the value 1 if a firm has previously filed a redacted agreement in our sample, prior to time t . Second, $\log(DaysSinceFirstRedact)$ captures the number of days since the first redacted agreement filed by the firm in question was filed. To the extent that we do not capture all redacted filings by a given firm, we assume that this measure captures prior redaction history assuming a uniform within-firm rate of redaction over time. Third, the variable $\log(\#PriorRedact)$ captures the number of prior redacted agreements filed by the given firm, prior to the filing of the current agreement. Fourth, the variable $RedactRate$ is simply the number of filings by the given firm which were redacted divided by the total number of filings (both redacted and un-redacted) submitted by the given firm, prior to the filing of the current agreement. The results are presented in Table 11. The coefficients on the interaction of each of these variables with $\mathbf{1}\{REDACT\}$ are significant and positive, suggesting that serial redactors experience worsened liquidity (measured by bid-ask spread) post-filing, when they redact, when compared with currently redacting firms which had not redacted as frequently in the past. Thus, the relationship between redaction and liquidity is concave, and firms can only credibly signal high value technology up to a point. Repeated, indiscriminate redaction seems to be viewed negatively by investors.

Third, we analyze whether a firm's degree of equity dependence affects the relationship between redaction and liquidity.⁶⁹ The argument is that firms which are more equity dependent have to turn to equity markets more frequently to raise financing, and thus are less subject to information asymmetry induced adverse selection problems. Thus, firms in more equity dependent industries should

⁶⁹We thank David Brown for this suggestion.

Table 3.17: Effect of Redaction Conditional on Equity Dependence

This table displays estimated regression coefficients and standard errors (in brackets) corresponding to the cross-sectional effects of firms' equity dependence. The regression takes the form $\Delta Y_{i,j,l,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,l,t} + \beta_2 Y_{i,j,l,t+k-1} + \beta_3 Z_{i,j,l,t} + \beta_4 \mathbf{1}\{REDACT\}_{i,j,l,t} \times Z_{i,j,l,t} + \mathbf{X}\gamma + \sum_l \alpha_l + \sum_t \alpha_t + \epsilon_{i,j,l,t}$. we use the Rajan and Zingales (1998) definition of industry level equity dependence, and create a dummy variable *HighDependence* which takes a value of 1 if a firm belongs to an industry in the 75th (or alternatively, 80th) percentile of industries ranked in increasing order of equity dependence. All other variables are as previously defined. The regressions are estimated by OLS with firm- and contract-level controls (specified in Appendix C.3), as well as technology-industry (SIC2) and year fixed-effects, with standard errors clustered at the technology-industry (SIC2) level.

| | $\Delta Illiq_{t+1}$ | $\Delta Illiq_{t+1}$ | $\Delta Spread_{t+1}$ | $\Delta Spread_{t+1}$ |
|--|----------------------|----------------------|-----------------------|-----------------------|
| $\mathbf{1}\{REDACT\}$ | -0.029*** [0.008] | -0.026*** [0.009] | -0.033*** [0.008] | -0.032*** [0.009] |
| $\mathbf{1}\{REDACT\} \times HighDependence$ | -0.026 [0.022] | -0.051** [0.023] | -0.063*** [0.023] | -0.077*** [0.025] |
| <i>HighDependence</i> | 0.003 [0.022] | 0.016 [0.026] | 0.045* [0.025] | 0.043 [0.028] |
| <i>Level_t</i> | -0.024*** [0.004] | -0.024*** [0.004] | -0.044*** [0.004] | -0.044*** [0.004] |
| Analysis | OLS | OLS | OLS | OLS |
| SIC2 + Year FE | Y | Y | Y | Y |
| Contract + Firm Controls | Y | Y | Y | Y |
| Equity Dependence Percentile | 75 th | 80 th | 75 th | 80 th |
| N | 3312 | 3312 | 3282 | 3282 |
| $\overline{R^2}$ | 0.254 | 0.255 | 0.262 | 0.262 |

face a greater positive effect of redaction on liquidity compared to firms which redact and belong to less equity dependent industries (who are more likely to be adversely selecting investors when they go to equity markets for financing). Accordingly, we use the Rajan and Zingales (1998) definition of industry-level equity financing dependence, and create a dummy variable *HighDependence* which takes a value of 1 if a firm belongs to an industry in the 75th (or alternatively, 80th) percentile of industries ranked in increasing order of equity dependence. The results in Table 12 show that the coefficients on the interaction term $\mathbf{1}\{REDACT\} \times HighDependence$ are negative and significant in three out of four cases. Thus, firms that are highly equity dependent and redact have an even greater improvement in liquidity (measured by bid-ask spread) post-filing, compared to firms which redact and belong to less equity dependent industries. This evidence suggests that

firms with greater equity dependence, who face the discipline of capital markets more frequently and are thus plausibly less affected by adverse selection problems, experience a positive marginal effect of redaction on stock liquidity.

3.7 Conclusion

In this chapter, we demonstrate that firms which redact information on the value of their IP in mandatory corporate disclosures experience positive capital market reactions as a result. We rationalize this finding with a stylized model in which non-disclosure serves as a credible signal of the value of a firm's technology. We demonstrate implications of this model in data. Using a novel licensing database, we show that strategic non-disclosure is rewarded, not penalized, by investors, evidenced through improved stock liquidity, increased equity issuance and institutional ownership. As a result of their ability to protect information on their IP, redacting firms subsequently increase their innovation activities relative to firms which fully disclose. In the cross-section, we find that signaling through redaction is not uniformly credible for all firms or at all times. Firms with lower reputations or which face greater information asymmetry, for example smaller or younger firms, experience a negative capital market reaction to their redaction decision.

This chapter opens the line for many avenues of future work. An immediate and obvious extension to this chapter would be to provide an extended theoretical framework to understand the total welfare implications of non-disclosure. Finding the right balance is of urgent need, with the emergence of new forms of financing for innovative firms, and the disclosure issue being central in the regulatory debates that surround these developments. Our results would suggest, however, that between the extremes of voluntary disclosure and mandatory disclosure – extremes which, at least in principle, favor the firm and the investor respectively – the appropriate balance is perhaps mandatory disclosure with partial exceptions.

REFERENCES

- Abadie, A. and J. Spiess (2016a). Robust Post-Matching Inference. *Working Paper*.
- Abadie, A. and J. Spiess (2016b). Robust post-matching inference. *Working Paper*.
- Admati, A. R. and P. Pfleiderer (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial studies* 1(1), 3–40.
- Agarwal, S., D. Lucca, A. Seru, and F. Trebbi (2014). Inconsistent Regulators: Evidence from Banking. *Quarterly Journal of Economics* 129(2), 889–938.
- Akcigit, U., M. A. Celik, and J. Greenwood (2016). Buy, keep, or sell: Economic growth and the market for ideas. *Econometrica* 84(3), 943–984.
- Ali, A., S. Klasa, and E. Yeung (2009). The limitations of industry concentration measures constructed with compustat data: Implications for finance research. *Review of Financial Studies* 22(10), 3839–3871.
- Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets* 5(1), 31–56.
- Anton, M. and C. Polk (2014). Connected stocks. *The Journal of Finance* 69(3), 1099–1127.
- Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance* 59(3), 1259–1294.
- Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609–626. Princeton University Press.
- Bailey, M., R. Cao, T. Kuchler, and J. Strobel (2016). Dp11272 social networks and housing markets.
- Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance* 61(4), 1645–1680.
- Barberis, N. and A. Shleifer (2003). Style investing. *Journal of financial Economics* 68(2), 161–199.

- Barberis, N., A. Shleifer, and J. Wurgler (2005). Comovement. *Journal of Financial Economics* 75(2), 283–317.
- Bartram, S. M., J. Griffin, T.-H. Lim, and D. T. Ng (2015). How important are foreign ownership linkages for international stock returns? *Review of Financial Studies*.
- Basak, S. and A. Pavlova (2013). Asset prices and institutional investors. *The American Economic Review* 103(5), 1728–1758.
- Ben-David, I., F. Franzoni, and R. Moussawi (2016). Exchange traded funds (etfs). Technical report, National Bureau of Economic Research.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004). How Much Should We Trust Differences-In-Differences Estimates? *Quarterly Journal of Economics* 119(1), 249–275.
- Bhattacharya, S. and J. R. Ritter (1983). Innovation and communication: Signalling with partial disclosure. *The Review of Economic Studies* 50(2), 331–346.
- Boone, A. L., I. V. Floros, and S. A. Johnson (2016). Redacting proprietary information at the initial public offering. *Journal of Financial Economics* 120(1), 102–123.
- Campello, M. and J. Gao (2016). Customer concentration and loan contract terms. *Journal of Financial Economics Forthcoming*.
- Chen, H., P. De, Y. J. 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.
- Chien, C. V., J. L. Contreras, C. A. Corrado, S. J. Graham, D. Hegde, A. K. Rai, and S. Vishnubhakat (2016). Comment to the sec in support of the enhanced disclosure of patent and technology license information. *Available at SSRN 2815618*.
- Choudhary, P., K. Merkley, and K. Schipper (2017). Direct measures of auditors' quantitative materiality judgments: Properties, determinants and consequences for audit characteristics and financial reporting reliability. *Working Paper*.
- Cookson, J. A. and M. Niessner (2016). Why dont we agree? evidence from a social network of investors. *Evidence from a Social Network of Investors (March 24, 2016)*.
- Da, Z. and X. Huang (2015). Harnessing the wisdom of crowds.
- Darrough, M. N. and N. M. Stoughton (1990). Financial disclosure policy in an entry game. *Journal of Accounting and Economics* 12(1-3), 219–243.

- Dimmock, S. G., W. C. Gerken, and N. P. Graham (2015). Is Fraud Contagious ? Career Networks and Fraud by Financial. *Working Paper*, 2015.
- Edmans, A., L. Goncalves-Pinto, Y. Wang, and M. Xu (2014). Strategic news releases in equity vesting months. Technical report, National Bureau of Economic Research.
- Egan, M., G. Matvos, and A. Seru (2016). The Market for Financial Adviser Misconduct. *Working Paper* (February), 1–60.
- Fama, E. F. and K. R. French (2016). Dissecting anomalies with a five-factor model. *Review of Financial Studies* 29(1), 69–103.
- Feltham, G. A. and J. Z. Xie (1992). Voluntary financial disclosure in an entry game with continua of types. *Contemporary Accounting Research* 9(1), 46–80.
- Foerster, S., B. T. Melzer, J. T. Linnainmaa, and A. Previtero (2014). Retail Financial Advice: Does One Size Fit All? *Working Paper*.
- Gans, J., S. Stern, and J. Wu (2016). The foundations of entrepreneurial strategy.
- Gans, J. S., D. H. Hsu, and S. Stern (2008). The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays. *Management Science* 54(5), 982–997.
- Gans, J. S. and S. Stern (2003). The product market and the market for 'ideas': Commercialization strategies for technology entrepreneurs. *Research Policy* 32(2), 333–350.
- Gertner, R., R. Gibbons, and D. Scharfstein (1988). Simultaneous signalling to the capital and product markets. *The RAND Journal of Economics*, 173–190.
- Giannini, R., P. Irvine, and T. Shu (2015). The convergence and divergence of investors opinions around earnings news: Evidence from a social network.
- Graham, J. R. and A. Kumar (2006). Do dividend clienteles exist? evidence on dividend preferences of retail investors. *The Journal of Finance* 61(3), 1305–1336.
- Green, T. C. and B.-H. Hwang (2009). Price-based return comovement. *Journal of Financial Economics* 93(1), 37–50.
- Greenwood, R. (2008). Excess comovement of stock returns: Evidence from cross-sectional variation in nikkei 225 weights. *Review of Financial Studies* 21(3), 1153–1186.
- Gurun, U. G., N. Stoffman, and S. E. Yonker (2015). Trust Busting : The Effect of Fraud on Investor Behavior. *Working Paper*.

- Hadlock, C. J. and J. R. Pierce (2010). New evidence on measuring financial constraints: Moving beyond the kz index. *Review of Financial Studies* 23(5), 1909–1940.
- Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). Market value and patent citations. *RAND Journal of Economics*, 16–38.
- Hall, B. H. and J. Lerner (2010). The financing of r&d and innovation. *Handbook of the Economics of Innovation* 1, 609–639.
- Hameed, A., R. Morck, J. Shen, and B. Yeung (2015). Information, analysts, and stock return comovement. *Review of Financial Studies*, hhv042.
- Hegde, D. (2014). Tacit knowledge and the structure of license contracts: Evidence from the biomedical industry. *Journal of Economics & Management Strategy* 23(3), 568–600.
- Hegde, D. and H. Luo (2017). Patent publication and the market for ideas. *Management Science*.
- Hoberg, G. and G. Phillips (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124(5), 1423–1465.
- Hoberg, G., G. Phillips, and N. Prabhala (2014). Product market threats, payouts, and financial flexibility. *The Journal of Finance* 69(1), 293–324.
- Holden, C. W. and S. Jacobsen (2014). Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. *The Journal of Finance* 69(4), 1747–1785.
- Hong, H., J. D. Kubik, and J. C. Stein (2004). Social interaction and stock-market participation. *The Journal of Finance* 59(1), 137–163.
- Imai, K., G. King, and E. Stuart (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171, 481–502.
- Israelsen, R. D. (2014). Does common analyst coverage explain excess comovement?
- Jiang, L., J. Liu, and B. Yang (2016). Communication and comovement: Evidence from online stock forums. *Available at SSRN* 2565250.
- Jung, M. J., J. P. Naughton, A. Tahoun, and C. Wang (2015). Corporate use of social media. *Available at SSRN*.
- K Pool, V., N. Stoffman, and S. E. Yonker (2014). The people in your neighborhood: Social interactions and mutual fund portfolios. *The Journal of Finance*.

- Kerr, W. R. and R. Nanda (2015). Financing innovation. *Annual Review of Financial Economics* 7, 445–462.
- Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2016). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics* Forthcoming.
- Kumar, A. and C. Lee (2006). Retail investor sentiment and return comovements. *The Journal of Finance* 61(5), 2451–2486.
- Kumar, A., J. K. Page, and O. G. Spalt (2013). Investor sentiment and return comovements: Evidence from stock splits and headquarters changes. *Review of Finance* 17(3), 921–953.
- Lerner, J. and A. Seru (2015). The use and misuse of patent data: Issues for corporate finance and beyond. *Booth/Harvard Business School Working Paper*.
- Llorente, G., R. Michaely, G. Saar, and J. Wang (2002). Dynamic volume-return relation of individual stocks. *Review of Financial studies* 15(4), 1005–1047.
- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies* 60(3), 531–542.
- Michaely, R., J. A. Popadak, and C. J. Vincent (2014). The deleveraging of us firms and institutional investors' role. *Available at SSRN 1941902*.
- Mitton, T., K. Vorkink, and I. J. Wright (2015). Neighborhood effects on speculative behavior. *Available at SSRN 2139106*.
- Peters, R. H. and L. A. Taylor (2016). Intangible capital and the investment-q relation. *Journal of Financial Economics*.
- Rajan, R. G. and L. Zingales (1998). Financial dependence and growth. *American Economic Review*, 559–586.
- Shapiro, C. (1985). Patent licensing and r & d rivalry. *The American Economic Review* 75(2), 25–30.
- Stambaugh, R. F. (2014). Presidential address: Investment noise and trends. *The Journal of Finance* 69(4), 1415–1453.
- Stock, J. H. and M. Yogo (2002). Testing for weak instruments in linear iv regression.
- Thakor, R. T. and A. W. Lo (2016). Optimal financing for r&d-intensive firms. *Working Paper*.

- Tumarkin, R. and R. F. Whitelaw (2001). News or noise? internet postings and stock prices. *Financial Analysts Journal* 57(3), 41–51.
- Varner, T. R. (2011). An economic perspective on patent licensing structure and provisions. *Business Economics* 46(4), 229–238.
- Veldkamp, L. L. (2006). Information markets and the comovement of asset prices. *The Review of Economic Studies* 73(3), 823–845.
- Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics* 5, 179–194.
- Verrecchia, R. E. and J. Weber (2006). Redacted disclosure. *Journal of Accounting Research* 44(4), 791–814.
- Wagenhofer, A. (1990). Voluntary disclosure with a strategic opponent. *Journal of Accounting and Economics* 12(4), 341–363.

A APPENDIX (CHAPTER 1)

Table A.1: Sample Selection Process

| | Subset | # |
|--|--------------|------------|
| Stocks existing between [2010-2015] | | 5458 |
| $N=5458 ; N*(N-1)/2$ | | 14894882 |
| CRSP SHRCD (10,11) Pairs That Ever Co-Existed | | 13732880 |
| Pairs that ever received 5 monthly coposts | | 1110234 |
| 22500 pairs randomly chosen from 16 size buckets | | 360000 |
| | Sample sizes | |
| Monthly (Sept2012-March 2016) ; 360,000 pairs, less attrition of stocks and controls | | 7,064,231 |
| Weekly (Sept2012-March 2016) ; 360,000 pairs, less attrition of stocks and controls | | 42,935,297 |

Table A.2: Variable Definitions

| Variable | Definition |
|--|---|
| $\log(1 + Coposts_{ijt})$ | For most analysis in this chapter, a single co-post is a poster-hour in which the poster discusses i or j , across all messages in the hour. In the instrumental variable analysis, I require that a co-post be, for certain reasons specific to that experiment, the mention of both legs in the stock pair in a single post. Across a month or week, it is the summation of all hours. A single poster can therefore contribute at most the number of hours in the month to this statistic. Hours are aligned so that if not during trading hours, it is aligned to the next trading day. |
| $\log(1 + CommonFollowers_{ijt})$ | For every hour, for every pair of posters ab where poster a discusses stock i in the hour, poster b stock j , I count the number of followers intersecting between a and b . The hourly value is the sum across all poster pairs; the monthly value is the sum of this value across hours. The followers are calculated from at least six months prior. In some of the analyses, I limit the observation window to 15 minutes or half hour instead of an hour. |
| $\log(1 + PairsCommonFollowers_{ijt})$ | For every hour, for every pair of posters ab where poster a discusses stock i in the hour, poster b stock j , I count the number of posters a and b with at least one intersecting follower. The hourly value is the sum across all poster pairs; the monthly value is the sum of this value across hours. The followers are calculated from at least six months prior. In some of the analyses, I limit the observation window to 15 minutes or half hour instead of an hour. |
| $\log(1 + AbovemedianCoowners_{ijt})$ | Based on Thomson Reuters 13-F, this is the number of owners that own both i and j the prior quarter, where their ownership of both stocks is above the quarter's cross-sectional median ownership for that stock. This is based on Gao, Ng and Moulton (2015) and Anton and Polk (2014). |
| ret_{it} | The return reported in CRSP for stock i compounded over the days that form the interval t . |
| $turnover_{it}$ | As reported in CRSP, this is the volume divided by shares outstanding. |

Table A.3: Average Annual Stock Characteristics

| Bin | # Msg. | PRC | Cap(10^9) | Value | 12-mo. ret | β_{MKT} | β_{SMB} | β_{HML} | σ | skewness | maxlretl |
|-----|-----------|---------|---------------|-------|------------|---------------|---------------|---------------|----------|----------|----------|
| 0 | 0 | 643.668 | 1.874 | 1.848 | 0.133 | 0.584 | 0.501 | 0.165 | 0.030 | 0.758 | 0.184 |
| 1 | 2.785 | 12.987 | 0.187 | 1.950 | 0.040 | 0.435 | 0.449 | 0.141 | 0.034 | 0.568 | 0.182 |
| 2 | 14.654 | 15.280 | 0.241 | 1.881 | 0.090 | 0.421 | 0.396 | 0.180 | 0.031 | 0.633 | 0.167 |
| 3 | 35.637 | 20.132 | 0.286 | 1.160 | 0.103 | 0.552 | 0.581 | 0.180 | 0.028 | 0.558 | 0.147 |
| 4 | 61.484 | 21.630 | 0.481 | 0.843 | 0.119 | 0.741 | 0.816 | 0.298 | 0.026 | 0.416 | 0.141 |
| 5 | 94.457 | 26.154 | 0.852 | 0.744 | 0.116 | 0.886 | 0.891 | 0.293 | 0.026 | 0.364 | 0.148 |
| 6 | 133.356 | 30.429 | 1.387 | 0.712 | 0.123 | 0.935 | 0.828 | 0.226 | 0.026 | 0.317 | 0.149 |
| 7 | 197.346 | 31.500 | 2.180 | 0.624 | 0.146 | 0.988 | 0.730 | 0.117 | 0.026 | 0.304 | 0.156 |
| 8 | 314.987 | 32.479 | 3.649 | 0.580 | 0.151 | 1.022 | 0.627 | 0.081 | 0.028 | 0.317 | 0.167 |
| 9 | 616.726 | 35.476 | 7.094 | 0.574 | 0.177 | 1.039 | 0.548 | 0.006 | 0.030 | 0.395 | 0.186 |
| 10 | 5,367.280 | 46.052 | 23.447 | 0.506 | 0.189 | 1.046 | 0.430 | -0.211 | 0.035 | 0.503 | 0.234 |

163

Table A.4: Most Popular Stocks in 2015

The fourth column is based on residuals of a cross-sectional regression of log posts on the stock return and characteristic variables in A.3, plus their squared terms.

| 1st tercile | 2nd tercile | 3rd tercile | Net of A Benchmark Linear Model |
|-------------------------------|-----------------------------|----------------------------------|---------------------------------|
| GREAT BASIN SCIENTIFIC INC | MANNKIND CORP | APPLE COMPUTER INC | IBIO INC |
| AMEDICA CORP | PLUG POWER INC | TWITTER INC | BIOCEPT INC |
| F X C M INC | SYNERGY PHARMACEUTICALS INC | GOPRO INC | MAGNEGAS CORP |
| SOLAR3D INC | KANDI TECHNOLOGIES CORP | NETFLIX INC | HEMISPHERX BIOPHARMA INC |
| GEVO INC | XOMA CORP | FACEBOOK INC | FUNCTIONX INC |
| BIOCEPT INC | GLU MOBILE INC | TESLA MOTORS INC | GOPRO INC |
| ASCENT SOLAR TECHNOLOGIES INC | PIPEX PHARMACEUTICALS INC | M E M C ELECTRONIC MATERIALS INC | F X C M INC |
| IBIO INC | LUMBER LIQUIDATORS INC | GILEAD SCIENCES INC | MANNKIND CORP |
| MAGNEGAS CORP | RELYPSA INC | FITBIT INC | SYNTHETIC BIOLOGICS INC |
| ACTINIUM PHARMACEUTICALS INC | ANAVEX LIFE SCIENCES CORP | AMAZON COM INC | PLUG POWER INC |

Table A.5: Additional summary statistics

In this table, I present correlations of other outcome variables alongside excess return correlation. The monthly outcome variables are derived from CRSP. Amihud illiquidity ratio is the absolute return over volume traded, assumed to be zero when no trading occurs. The weekly variables are based on DTAQ and the correlations are computed at the weekly level across half-hour intervals.

Panel A: Monthly

| | μ | σ | 1st | 5th | 25th | 40th | 50th | 60th | 75th | 99th |
|-------------------------------|-------|----------|--------|--------|--------|-------|-------|-------|-------|-------|
| $\rho_{ijt}^{ret^\epsilon}$ | 0.93 | 24.09 | -53.96 | -38.21 | -15.39 | -5.33 | 0.74 | 6.84 | 17.03 | 57.94 |
| ρ_{ijt}^{ret} | 24.53 | 26.46 | -39.38 | -20.39 | 6.59 | 18.40 | 25.48 | 32.44 | 43.67 | 78.53 |
| $\rho_{ijt}^{ ret^\epsilon }$ | 1.88 | 23.55 | -44.06 | -32.48 | -14.99 | -6.23 | -0.46 | 5.67 | 16.61 | 64.65 |
| ρ_{ijt}^{Amihud} | 8.83 | 24.52 | -43.69 | -30.46 | -8.82 | 1.69 | 8.24 | 14.89 | 25.97 | 64.88 |
| $\rho_{ijt}^{turnover}$ | 13.40 | 28.24 | -41.99 | -28.97 | -7.86 | 3.47 | 11.06 | 19.10 | 33.01 | 80.03 |
| ρ_{ijt}^{Volume} | 12.79 | 28.08 | -42.12 | -29.21 | -8.31 | 2.87 | 10.38 | 18.33 | 32.15 | 79.67 |
| $\rho_{ijt}^{\%Spread}$ | 3.75 | 27.17 | -54.82 | -35.51 | -14.50 | -5.54 | 0.87 | 7.82 | 19.89 | 81.47 |

Panel B: Weekly

| | μ | σ | 1st | 5th | 25th | 40th | 50th | 60th | 75th | 99th |
|-------------------------------|-------|----------|--------|--------|-------|-------|-------|-------|-------|-------|
| ρ_{ijt}^{ret} | 16.12 | 21.79 | -39.28 | -17.86 | 2.06 | 10.03 | 15.32 | 20.84 | 30.36 | 67.05 |
| $\rho_{ijt}^{ret^\epsilon}$ | 3.06 | 20.12 | -46.76 | -28.83 | -9.08 | -1.68 | 2.61 | 7.04 | 14.92 | 56.05 |
| $\rho_{ijt}^{ ret^\epsilon }$ | 14.41 | 19.23 | -19.08 | -12.10 | 0.18 | 7.23 | 11.95 | 16.97 | 25.93 | 69.87 |
| $\rho_{ijt}^{OrderImbal.}$ | 1.56 | 22.22 | -67.62 | -31.58 | -9.46 | -2.63 | 1.25 | 5.18 | 12.29 | 72.89 |
| $\rho_{ijt}^{ OrderImbal. }$ | 13.91 | 21.26 | -17.19 | -10.66 | -0.80 | 5.01 | 9.14 | 13.81 | 22.99 | 88.00 |
| ρ_{ijt}^{Amihud} | 25.92 | 16.61 | -10.08 | -0.26 | 14.34 | 21.15 | 25.38 | 29.69 | 36.98 | 66.20 |
| ρ_{ijt}^{Spread} | 61.64 | 24.10 | -11.18 | 8.57 | 50.68 | 61.71 | 67.19 | 71.99 | 78.77 | 97.31 |
| ρ_{ijt}^{Volume} | 39.09 | 19.79 | -5.23 | 5.55 | 25.05 | 34.22 | 39.71 | 45.08 | 53.58 | 80.47 |

Table A.6: User-level coverage statistics

This table presents statistics describing the number of stocks, sectors or industries covered by a single user throughout their entire posting history. *VW* and *EW* indicate value and equal weighting, where value-weighting refers to the number of connections the user has in the followership network as of August 2016, accounting for the tendency of some users to appear more in the "common follower" statistic than others. Sectors and industries are defined according to StockTwits. There are roughly 50 sectors and 200 industries.

| | # Symbols | #Sectors | #Industry |
|----------------|-----------|----------|-----------|
| Value-weighted | 301.72 | 8.90 | 123.29 |
| Equal-weighted | 41.42 | 4.46 | 15.20 |

This table quantifies the extent to which mentions of a same stock, industry or sector persist. For a single 3 month period, I calculate the probability that same stock will be mentioned 3 quarters later.

| %Firm | % Industry | %Sector |
|-------|------------|---------|
| 17.54 | 22.92 | 28.25 |

This table quantifies the probability a user will mention the same stock ever again until the end of the sample period, March 2016. For a given year, say 2012, the probability is based on all stocks the user mentioned prior to the January of that year.

| Year | % Probability |
|------|---------------|
| 2012 | 26.07 |
| 2013 | 19.45 |
| 2014 | 16.61 |
| 2015 | 14.27 |

Table A.7: Timing decomposition

This table decomposes postings at the monthly level into halves or quarters. The hypothesis is that earlier posts should affect comovement more than later posts, if social media audiences react to postings with a delay.

| | Halves | Halves Z-scored | Quarters | Quarters Z-scored |
|-------------------------------------|---------------------|---------------------|---------------------|----------------------|
| $\log(1+\text{Coposts}_{ijt})_{H1}$ | 0.244*** (0.032) | 0.298*** (0.039) | | |
| $\log(1+\text{Coposts}_{ijt})_{H2}$ | 0.172*** (0.031) | 0.206*** (0.038) | | |
| $\log(1+\text{Coposts}_{ijt})_{Q1}$ | | | 0.242*** (0.043) | 0.108*** (0.019) |
| $\log(1+\text{Coposts}_{ijt})_{Q2}$ | | | 0.342*** (0.058) | 0.172*** (0.029) |
| $\log(1+\text{Coposts}_{ijt})_{Q3}$ | | | 0.256*** (0.056) | 0.131*** (0.029) |
| $\log(1+\text{Coposts}_{ijt})_{Q4}$ | | | 0.144*** (0.041) | 0.066*** (0.019) |
| Num. obs. | 7,060,175 | 7,060,175 | 7,060,175 | 7,060,175 |
| FE? | Y | Y | Y | Y |
| Controls? | Y | Y | Y | Y |
| R ² | 0.074 | 0.074 | 0.074 | 0.074 |
| Adj. R ² | 0.051 | 0.051 | 0.051 | 0.051 |

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The screenshot shows the MarketWatch website interface. At the top, there's a navigation bar with 'Log In', 'MarketWatch', and 'Virtual Stock Exchange'. The main content area features a large headline: 'Biotech gets the blame as Monday's rally unravels'. Below the headline is a photo of Hillary Clinton speaking at a podium. To the right of the photo are several sub-headlines related to the biotech sector and stock market. A 'Markets' sidebar on the right displays a table of market indices and a line chart. At the bottom, there's a 'Trending Tickers' section and a 'StockTwits' banner.

Latest News

- 1:37p Oil rebounds on expectations for further output declines
- 1:21p U.S. stock rally fizzles out, led by biotech selloff
- 1:18p Let the big picture keep your portfolio safe
- 1:18p China's clumsy meddling now threatens Hong Kong's success
- 4:00p Dollar declines as Fed still seen as first central bank to

Virtual Stock Exchange

Our free stock-market game

- Trade your virtual portfolio in real time
- Talk strategies in group discussions
- Find or create a game that suits you
- Use our learning center to improve

BULLETIN Health-care stocks play central role in unraveling of rally »

Share this bulletin
Get news bulletins by email [Tweet](#) X

Home News Viewer Markets Investing Trading Deck Personal Finance Retirement Economy Real Estate Portfolio Watchlist Alerts Games

HOT TOPICS · Location Scouts · Pope Francis · Climate · MARKETWATCH VIDEO · Turning 50 · HOT STOCKS · NFLX · AMZN · P

Biotech gets the blame as Monday's rally unravels

Health-care sector takes it on the chin as stock rally unravels

Hillary Clinton says she'll lay out plan against gouging by specialty-drug makers

Not just Hillary Clinton who thinks drug prices are insane

We fret over finances more than health

Getty Images

Volkswagen loses €14 bln **NEED TO KNOW**
U.S. existing-home sales drop nearly 5%

Markets

| | PRICE | CHG | %CHG | RANGE: | 1 DAY |
|-----------|--------|-------|-------|------------|--------|
| Dow | 16,486 | +103 | 0.63% | | 16,579 |
| Nasdaq | 4,833 | +6 | 0.12% | | 16,486 |
| S&P 500 | 1,967 | +9 | 0.46% | | |
| GlobalDow | 2,319 | -7 | 0.32% | | |
| Gold | 1,133 | -5 | 0.42% | | 16,392 |
| Oil | 46.30 | +1.62 | 3.63% | 10a 12p 2p | |

Most Popular

- NEED TO KNOW**
Gold prices are set to jump, this pattern suggests
- MARKET SNAPSHOT**
U.S. stock rally fizzles out, led by biotech selloff

Trending Tickers IBB -4.16% GILD -2.64% XBI -6.01% BLUE -10.39% XON -5.17%

Powered by **StockTwits** X

Figure A.1: StockTwits integration with Yahoo! Finance and Marketwatch

StockTwits is integrated into other platforms, such as Marketwatch.com, one of the world's most popular financial news media sites. Integration has been in place since September 2014. The StockTwits banner is at the bottom of the page.

Yahoo! Inc. (YHOO) 1:33pm EDT: **31.06** ↑ **0.32 (1.04%)**

More On YHOO

QUOTES

Summary

Order Book

Options

Historical Prices

CHARTS

Interactive

NEWS & INFO

Headlines

Press Releases

Company Events

Message Boards

▶ **Market Pulse**

COMPANY

Profile

Key Statistics

SEC Filings

Competitors

Industry

Components

ANALYST COVERAGE

Analyst Opinion

Analyst Estimates

OWNERSHIP

Major Holders

Insider Transactions

Insider Roster


FINANCIALS


Income Statement


Balance Sheet


Cash Flow


Market Pulse for YHOO


 **WealthMgt2010**
 \$YHOO Damn, what a reversal...anybody that bought this morning is now underwater.
ST 56 minutes ago

 **master_trader42**
 \$BABA down but \$YHOO up? does this mean YHOO bottomed?
ST 1 hour 35 minutes ago

 **1986iamwallstreet**
 Alibaba s \$105 Billion Lockup Ends, Putting Focus on Yahoo Stake
<http://www.bloomberg.com/news/articles/2015-09-20/alibaba-s-105-billion-lockup-ends-putting-focus-on-yahoo-stake> \$YHOO \$BABA
ST 1 hour 38 minutes ago

 **TheStreet**
 Here s the Reason Yahoo! Isn t Touting Its Own Daily Fantasy Sports \$YHOO
http://www.thestreet.com/story/13295250/1/here-s-the-reason-yahoo-isn-t-touting-its-own-daily-fantasy-sports.html?pucc=stocktwits&cm_ven=STOCKTWITS&utm_source=dvr.it&utm_med
ST 1 hour 58 minutes ago

 **WealthMgt2010**
 @VincentVG: \$YHOO Hopefully \$YHOO selling it s stake in \$BABA today...lololololol What s so funny..\$BABA has been selling off all day
ST 2 hours 1 minute ago

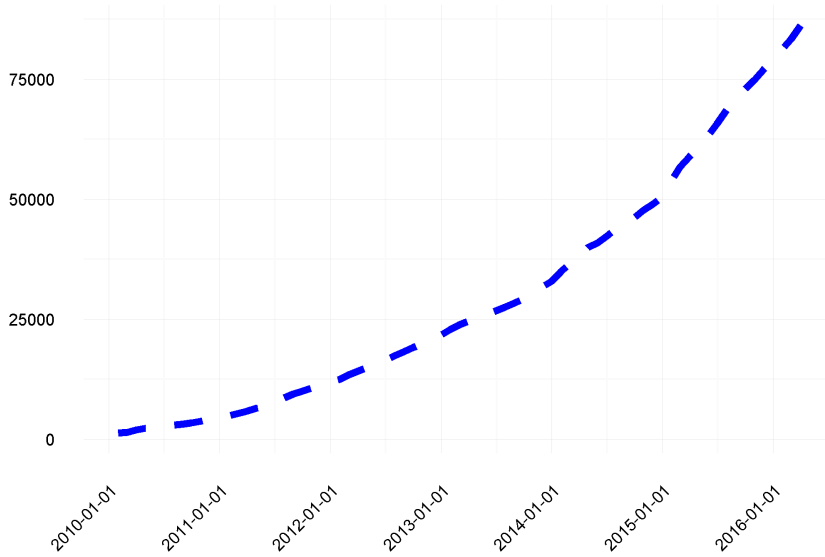
 **VincentVG**
 @WealthMgt2010: \$YHOO Hopefully \$YHOO selling it s stake in \$BABA today...lololololol
ST 2 hours 30 minutes ago

Get Market Pulse for:

Trending Tickers on Market Pulse

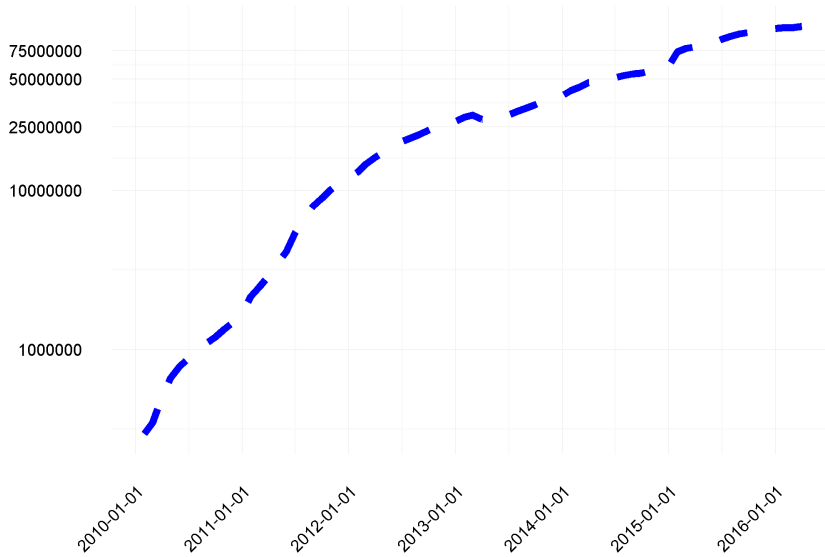
| Name | Symbol |
|---------------------------------|--------|
| Amerco | UHAL |
| iShares Nasdaq Biotechnology In | IBB |
| WaferGen Bio-systems, Inc. | WGBS |
| Express Scripts Holding Company | ESRX |
| SPDR Series Trust SPDR S&P Bio | XBI |
| bluebird bio, Inc. | BLUE |
| Valeant Pharmaceuticals Interna | VRX |
| ProShares Ultra Nasdaq Biotechn | BIB |
| Direxion Daily S&P Biotech Bull | LABU |
| Calithera Biosciences, Inc. | CALA |

This page examples another major StockTwits partner (during the sample period) was Yahoo! Finance.



Panel A: The number of StockTwits users with at least 50 followers

In this graph, I plot the number cumulative posters across StockTwits with at least 50 posts as of their latest post prior to the indicated month.



Panel B: The number of followers-in-common across posters

In this graph, I plot the number of followers common to all pairs of posters with at least 50 posts as of their latest post prior to the indicated month.

Figure A.2: Follower network over time

B APPENDIX (CHAPTER 2)

Table B.1: State Security Regulators

The table below shows the names of state securities regulators, broken up into divisions and departments. Divisions are the smallest organizational entities that oversee securities regulation.

For most states, the securities division is a sub-organization of a larger department like the Secretary of State or the Attorney General's office. The value is blank when a departmental hierarchy is not provided.

| State | Division Name | Department Name |
|----------------|---|--|
| Alabama | Alabama Securities Commission | |
| Alaska | Banking and Securities Division | Department of Commerce, Community, and Economic Development |
| Arizona | Securities Division | Arizona Corporation Commission |
| Arkansas | Arkansas Securities Department | |
| California | Securities Regulation Division | Department of Business Oversight |
| Colorado | Division of Securities | Department of Regulatory Agencies |
| Connecticut | Securities and Business Investment Division | Department of Banking |
| Delaware | Investor Protection Unit | Attorney General |
| Florida | Division of Securities | Office of Financial Regulation |
| Georgia | Division of Securities | Secretary of State Office |
| Hawaii | Division of Securities | Department of Commerce and Consumer Affairs |
| Idaho | Securities Section | Department of Finance |
| Illinois | Securities Department | Secretary of State |
| Indiana | Securities Division | Secretary of State |
| Iowa | Securities Bureau | Insurance Division |
| Kansas | Office of the Securities Commissioner | |
| Kentucky | Securities Division | Department of Financial Institutions |
| Louisiana | Securities Division | Office of Financial Institutions |
| Maine | Office of Securities | Department of Professional and Financial Regulation |
| Maryland | Securities Division | Attorney General |
| Massachusetts | Securities Division | Secretary of Commonwealth |
| Michigan | Corporations, Securities, and Commercial Licensing Bureau | Department of Licensing and Regulatory Affairs |
| Minnesota | Securities, Franchises, and Subdivided Land | Department of Commerce |
| Mississippi | Securities Division | Secretary of State |
| Missouri | Securities Division | Secretary of State |
| Montana | Commissioner of Securities and Insurance | Office of the Montana State Auditor |
| Nebraska | Department of Banking and Finance | |
| Nevada | Nevada Securities Center | Secretary of State |
| New Hampshire | Bureau of Securities Regulation | Secretary of State |
| New Jersey | Bureau of Securities | Division of Consumer Affairs |
| New Mexico | Securities Division | Regulation and Licensing Department |
| New York | Investor Protection Bureau | Attorney General |
| North Carolina | The Securities Division | Secretary of State |

Table B.2: State Security Regulators (continued)

| State | Division Name | Department Name |
|----------------|---|--------------------------------------|
| North Dakota | Securities Department | |
| Ohio | Division of Securities | Department of Commerce |
| Oklahoma | Department of Securities | |
| Oregon | Division of Financial Regulation | |
| Pennsylvania | Department of Banking and Securities | |
| Rhode Island | Department of Business Regulations | |
| South Carolina | Securities Division | Attorney General |
| South Dakota | Division of Securities | Division of Licensing and Regulation |
| Tennessee | Department of Commerce & Insurance | Department of Commerce & Insurance |
| Texas | State Securities Board | |
| Utah | Division of Securities | |
| Vermont | Securities Division | Department of Financial Regulation |
| Virginia | Division of Securities and Retail Franchising | State Corporation Commission |
| Washington | Division of Securities | Department of Financial Institutions |
| West Virginia | Securities Commission | State Auditor's Office |
| Wisconsin | Division of Securities | Department of Financial Institutions |
| Wyoming | Investing Center | Secretary of State |

B.1 Redacting Customer Complaints

One potential concern regarding our results is that redactions rose for firms that were treated relative to those that were not treated, driving our results. First, it seems plausible that state regulators may be more susceptible to regulatory capture, seeking to preserve the presence of investment advisers who may otherwise leave the state. Second, if effect is driven by the inability of adviser representatives to redact their complaints, being in a state with more regulatory staff means there are more resources to process the complaints. Moreover, being further away from the corresponding state regulator would also mean that redacting a complaint is more costly. We argue both of these is not likely in light of the legal environment through which redactions are processed. We discuss this in detail below.

Records are stored at FINRA through the Central Repository Deposit (CRD) system. Complaints that have alleged damages over \$5,000 or resulted in some legal action are both reported in the system. Upon receiving a complaint, both the investment adviser firm and investment adviser

representative have to file a complaint disclosure to the CRD system. In August 2010, FINRA began disclosing all historic complaints, regardless of age. In the past, unproven allegations were not disclosed after two years. Specifically, there are two types of complaints that investment adviser representatives may want to remove over which advisers have little control:

1. **Denied customer complaints.** Although denied customer complaints may seem insignificant, accusations typically are accompanied by harsh words that remain on the CRD for at least two years (since 2009, this stays on for 10 years). Even if an adviser's record shows patterns of denied rather than arbitrated or settled complaints, firms also may be hesitant to affiliate with that adviser. Moreover, whether or not complaints are settled in the first place is mainly up to the investment adviser firm, not the representative him/herself.
2. **Termination explanations.** Broker-dealers may terminate advisers for any reason. There may be discrepancies between the self-reported termination explanation and the firm-reported explanation.

Investment advisers occasionally request for expungement of customer complaints. Since the records are stored in the CRD, all expungement requests are handled by FINRA. Nonetheless, other regulators are involved in the process. FINRA may agree to remove disclosures if brokers obtain a recommendation that is false, erroneous, or that the broker wasn't involved in the alleged misdeed. To do this, representatives must acquire a court confirmation after submitting the expungement request. Once the request is submitted, the corresponding investment adviser regulator (SEC or state regulator) is informed, giving them a chance to oppose the expungement. State regulators received a total of 519 requests in 2010, up from 110 in 2009. In total, the process to expunge a complaint typically takes at least one year. Although FINRA claims to have tracked the number of expungements granted, they do not publicly disclose it.

Although the expungement process is fairly difficult, FINRA arbitrations could be settled subject to an agreement that claimants would not oppose the investment adviser representative's subsequent efforts to seek expungement from a court of competent jurisdiction. Subsequently, rep-

representatives would initiate unopposed petitions for expungement in state courts that were often rubber-stamped. The judge's order would then be submitted to FINRA, and the arbitration disclosure would be expunged. In response to this practice, FINRA adopted Rule 2130 in 2004. One of the most significant changes was the need to name FINRA as an additional party challenging party to expungement. This meant that FINRA also receives all appropriate documents with expungement, unambiguously increasing the cost of expungement requests. Moreover, although expungement requests from arbitrated cases are mostly granted, less than 8% of disclosures are expunged. Of the 7,621 arbitration cases from 2012 to 2014, only 563 records were expunged, according to the arbitration bar association.

Some of the surge in requests is also the result of new disclosure demands by FINRA. Until 2009, only brokers who were named as a party to a case had to disclose a customer complaint. Because most investors sue only the brokerage firm, that left a lot of accused individual brokers with clean records despite complaints that they had mishandled an account. Prior to this, large brokerage firms could shield individual brokers. However, it is not clear whether they will have this incentive. Larger firms may also be more likely to place blame on an individual whom they could terminate, in order to shift blame to the individual. After 2009, FINRA modified this disclosure practice, requiring all brokers to report complaints regardless of whether they were named directly as a respondent.

The institutional setting suggests that deletions of customer complaints is not relevant for the timing of the Dodd-Frank Act. Our phone calls with 3 regulators, Maryland, the SEC, and California also suggest that expungement is not an issue. Moreover, the censoring bias from any deleted complaints should not be correlated to treatment either. Finally, our specifications with state-year fixed effects and firm fixed effects absorb a lot of the drivers of expungement.

C APPENDIX (CHAPTER 3)

C.1 Full Model Analysis

In this appendix, we solve the model presented in Section 3.3. In order to do this, we consider the four possible strategies available to A . The first strategy is $\{1, 1\}$ where A discloses \tilde{t} regardless of its type. In this case, B enters only when it observes $\tilde{t} = \bar{t}$. Even though B does not incur the search cost, it is still not worth entering as $\frac{t}{2} - c < 0$. So, A receives \underline{t} . When $\tilde{t} = \bar{t}$, B enters and receives (in expectation) $\frac{\bar{t}}{2} - c$ and A receives $\frac{\bar{t}}{2}$. Thus, the payoffs for A to this strategy are $\{\underline{t}, \frac{\bar{t}}{2}\}$. The second strategy is $\{0, 1\}$, where A discloses if its type is high $\tilde{t} = \bar{t}$, but not if its type is low. Again, in this case B only enters when it observes $\tilde{t} = \bar{t}$. B does not enter when $D = 0$ as this indicates $t = \underline{t}$ with probability 1. This is because B 's potential expected profit from entry is $\frac{t}{2} - c - \underline{s} < 0$, so it does not enter. So, A 's profit is \underline{t} . When B observes $\tilde{t} = \bar{t}$, it enters and receives (in expectation) $\frac{\bar{t}}{2} - c$ and A receives $\frac{\bar{t}}{2}$. A 's payoffs are thus $\{\underline{t}, \frac{\bar{t}}{2}\}$.

The next strategy is $\{1, 0\}$, where A discloses if its type is low, $\tilde{t} = \underline{t}$, but not if its type is high. In this case, B does not enter regardless of type. The rents in the low-type market are not attractive enough, as $\frac{t}{2} < c$, and the search cost is prohibitively high in the high-type market, as $\frac{\bar{t}}{2} < \bar{s}$. Thus, A 's profits are $\{\underline{t}, \bar{t}\}$. The final strategy is $\{0, 0\}$. In this case, firm B 's belief on A 's type is $t' = \theta\bar{t} + (1 - \theta)\underline{t}$. Since $\underline{s} < s(t') = s' < \bar{s}$, we require:

$$\theta > \underline{\theta} = \frac{2(c + s') - \underline{t}}{\bar{t} - \underline{t}}$$

such that B enters. This imposes an lower bound on the probability of A 's technology being high-type, $\theta > \underline{\theta}$. This condition is more likely to hold if, first, $(\bar{t} - \underline{t})$ is large, or, second, if the entry and search costs are not too high relative to \underline{t} . Specifically, since θ is a probability, we need $\underline{\theta} \geq 0$. This implies that we need $c + s' \geq \frac{t}{2}$. We know that $\frac{t}{2} - c < 0$ by assumption, and $s' > 0$, so, $s' > \frac{t}{2} - c \implies c + s' > \frac{t}{2}$ as required. If $\theta > \underline{\theta}$, B enters and A 's expected payoffs are $\{\frac{t}{2}, \frac{\bar{t}}{2}\}$. We

summarize the payoffs of the game in the following table:

| | | | |
|-----------------------|---------|--|--|
| | | $\tilde{t} = \underline{t}$ | |
| | | $D = 1$ | $D = 0$ |
| $\tilde{t} = \bar{t}$ | $D = 1$ | $\{\frac{\bar{t}}{2}, \underline{t}\}$ | $\{\frac{\bar{t}}{2}, \underline{t}\}$ |
| | $D = 0$ | $\{\bar{t}, \underline{t}\}$ | $\{\frac{\bar{t}}{2}, \frac{t}{2}\}$ |

There are two Nash equilibria in pure strategies, $\{0, 1\}$ and $\{1, 0\}$, however only $\{1, 0\}$ survives iterated elimination of weakly dominated strategies.

C.2 Sample Construction and Selection

In this appendix, we provide details on the sample construction and selection procedure.

We first concord the data to financial databases, then disambiguate the economic role of the filer in the transaction (i.e. classify the filer as the licensor or the licensee). Finally, we filter observations according to our sample requirements. As aforementioned, table 3.2 reports the sample sizes at each stage of the filtering process.

To concord the data with financial databases, we first use RoyaltyStat’s provided *gvkeys*, followed by SEC *CIK* codes and filer names. RoyaltyStat maintains filer names and *CIK* codes as reported in the EDGAR log, and the point-in-time *gvkey*. The remainder of firms, for which such identifiers are unavailable, are firms that RoyaltyStat could not link to Compustat due to omissions in their coverage, or because while they do have publicly filed securities, they are not covered by Compustat. The company suggests that this is most often the case with penny stocks.⁷⁰ This translates to about 8,500 contracts with a filer *gvkey* referring to a firm in Compustat North America.

Second, we augment RoyaltyStat’s coverage of filers using the filer name and the filer *CIK*. Given a filer *CIK*, we coalesce possible names for the company using point-in-time *CIK* and the Capital IQ “Helper” table available on WRDS. Given a filer name, we standardize frequently occurring business words (“company” to “co”, “limited” to “ltd”) and do exact name matching on company names coalesced from Compustat, CRSP, Capital IQ and other financial databases that we have at our disposal. This yields approximately 1,000 additional matches for the identifiers not covered by RoyaltyStat.

Third, we concord each contract to the identifier of the *filer* and then identify whether the filer was the licensor or the licensee in the transaction. In order to ensure data integrity, we keep only contracts in which the filer can clearly be disambiguated as a licensor or licensee in the

⁷⁰We are in efforts to expand our data to this set, but this draft only includes firms that can be concorded to a firm in Compustat. In many studies, the excluded firms including penny stocks would be infra-marginal, but when discussing the role of disclosure for young, innovative firms, this set of firms may be especially important.

transaction. To do this disambiguation, we first match the filer with the name, *CIK* or *gvkey* provided by RoyaltyStat for the licensor or licensee. Through RoyaltyStat-provided data of 9,678 records, we are initially unable to assign the filer as the licensor or licensee in 2,824 cases. These remainder cases are either due to ambiguity, or because RoyaltyStat's efforts on identifying the licensor and licensee are to this date still in progress.

To match the remainder records, RoyaltyStat provided us a list of company names in its database. We first apply programmatic rules to match the transaction parties to the filers and then hand-match the remainder.⁷¹

Fourth, we exclude contracts that have a per-unit royalty, which RoyaltyStat flagged as a related party transaction, or which we flag as ambiguous. The final result set has 7,972 observations, though sample size varies across analyses due to vastly different data coverage across the various data sources that we use.

⁷¹We process list programmatically by doing three things. First, we flag the counter-party as an individual, university or government institute; in these cases, the filer must be the opposite party. Second, we write down a list of nearly 1,500 stem words, such that if the licensor contains the stem word and the filer contained the stem word, it is very unlikely for the filer to be the licensee, and vice versa. Third, we do exact name matching the subsidiaries of the firm based on CorpWatch. Finally, for the remainder cases, RoyaltyStat allowed us to export a list of filer names and we disambiguated (where possible) the role of the filer in relation to the economic transaction when possible. For 1,061 of 2,824, we were not (yet) able to match the filer to the licensor or the licensee. A substantial fraction of these cases are due to ambiguity and limited information particularly for transactions occurring in some cases over twenty years ago. We plan on continuing to augment this list.

C.3 Control Variable Definitions

The firm-level controls are the logarithm of total assets ($\log(\text{Assets})$), the logarithm of total intangible capital as defined by Peters and Taylor (2016) ($\log(K_{int})$), the ratio of cash to assets ($\frac{\text{Cash}}{\text{Assets}}$), leverage ratio ($\frac{\text{Debt}}{\text{Assets}}$), and the *SmallYoung* index value which collapses the age and size of the firm into a single index. The contract-level controls are dummy variables which take the value of 1 if the filing firm is the licensor (*FilerLicensor*), if the licensor is an individual (*LicensorIsIndividual*), a non-profit (*LicensorIsNonProfit*), a university (*LicensorIsUniversity*), if the given contract is an amendment of a previous agreement (*IsAmendment*), if the underlying IP consists of know-how (*IsKnowHow*), and show-how (*IsShowHow*). We also include the logarithm of the duration of the contract ($\log(\text{Duration})$), a dummy for whether the licensing rights extend worldwide (*Worldwide*), whether the agreement confers the licensee the right to sublicense the IP (*Sublicense*), and if the IP is being licensed exclusively (*Exclusive*).

C.4 Robustness Tests

Table C.1: Effect of Redaction on Future Liquidity in Event-Time

This table displays regression coefficients (and standard errors in brackets) of weekly liquidity measures around filing events. The liquidity variables are standardized by the event mean and variance, and therefore can be interpreted in standard deviations. The variable *EffSpread* is the weekly effective spread defined as twice the absolute value of the difference between the actual trade price and the midpoint of the market quote (i.e., between the quoted bid price and the quoted ask price), divided by the midpoint between these two prices based on DTAQ. The variable *PrcImpact* is the weekly price impact. The values are the weekly median of daily values to minimize the influence of outliers. *Post* is a variable that takes the value of 1 for all observations that fall in the 2 weeks or 15 weeks (depending on event window being considered) immediately following the filing date of the disclosure, and 0 otherwise. The standard errors are clustered by event.

| | <i>EffSpread</i> | <i>EffSpread</i> | <i>PrcImpact</i> | <i>PrcImpact</i> |
|------------------------------------|----------------------|----------------------|----------------------|----------------------|
| Event window | 15 weeks | 2 weeks | 15 weeks | 2 weeks |
| $\mathbf{1}\{REDACT\}$ | 0.064*** [0.018] | 0.066** [0.029] | 0.029*** [0.009] | 0.046* [0.025] |
| $\mathbf{1}\{REDACT\} \times Post$ | -0.114*** [0.032] | -0.057*** [0.020] | -0.052*** [0.016] | -0.061*** [0.020] |
| <i>Post</i> | -0.126*** [0.025] | -0.024 [0.015] | -0.013 [0.016] | -0.012 [0.020] |
| Constant | 0.070*** [0.013] | -0.031** [0.013] | 0.007 [0.009] | -0.009 [0.014] |
| N | 109452 | 15945 | 106025 | 15564 |
| $\overline{R^2}$ | 0.008 | 0.001 | 0.000 | 0.009 |

Table C.2: Robustness with SIC2-by-Year Fixed-Effects

This table displays results of specifications with industry-by-year fixed-effects. The regression takes the form $\Delta Y_{i,j,t,t+k} = \beta_1 \mathbf{1}\{REDACT\}_{i,j,t,t} + \beta_2 Y_{i,j,t,t+k-1} + \mathbf{X}\gamma + \sum_l \alpha_l \times \alpha_t + \epsilon_{i,j,t,t}$. All variables are as previously defined. We include industry (SIC2 of the underlying technology)-by-year fixed-effects. Standard errors for the OLS analysis (in brackets) are clustered at the technology industry (SIC2) level.

| | $\Delta \log(TSM)_{t+1}$ | $\Delta \log(TSM)_{t+2}$ | $\Delta Spread_{t+1}$ | $\Delta PrcImpact_{t+1}$ | $\Delta InstOwn_{t+1}$ | $\mathbf{1}\{Issuance\}_{t+1}$ |
|-----------------------------|--------------------------|--------------------------|-----------------------|--------------------------|------------------------|--------------------------------|
| $\mathbf{1}\{REDACT\}$ | 0.404*** [0.088] | 0.295** [0.116] | -0.017** [0.007] | -0.042*** [0.009] | 0.480*** [0.050] | 0.044*** [0.017] |
| $Level_t$ | | | -0.203*** [0.016] | -0.046*** [0.004] | | |
| $\log(TotalPriorPatents)_t$ | 0.420*** [0.028] | 0.400*** [0.026] | | | | |
| Analysis | OLS | OLS | OLS | OLS | OLS | OLS |
| SIC2-by-Year FE | Y | Y | Y | Y | Y | Y |
| Contract + Firm Controls | Y | Y | Y | Y | Y | Y |
| N | 3316 | 2979 | 3324 | 3303 | 4674 | 4110 |
| $\overline{R^2}$ | 0.533 | 0.505 | 0.290 | 0.311 | 0.333 | 0.211 |