ESSAYS IN ASSET PRICING AND INVESTOR BEHAVIOR

By

Qian Yang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Business Administration—Finance—Doctor of Philosophy

2023

ABSTRACT

In Chapter One, we examine the following question. Have retail investors become the ants that move the log? Social media has proved instrumental for effective coordination that might lead to extreme returns. To study this effect, I construct a novel crash risk measure by estimating ex-ante crash probabilities via logit and machine learning techniques. Stocks with high ex-ante crash risk tend to have lower returns, especially when lagged sentiment is high. Robinhood traders tend to over-buy high crash-risk stocks, consistent with the optimal expectations theory. By exploiting the staggered first appearances of ticker names on "Wallstreetbets", I document a causal effect of social transmission on crash risk. This effect is significantly more substantial for smaller stocks. To further bolster the finding, I exploit the entire history of Reddit to construct a novel instrument and show that social transmission is likely to cause elevated crash risk on a daily basis.

In Chapter Two, we examine the following issue. Cyber risk is an important but latent source of risk in the economy. To estimate its impact on the asset market, we use machine learning techniques to develop a firm-level measure of cyber risk. The measure aggregates information from a rich set of firm characteristics and shows superior ability to forecast future cyberattacks on individual firms. We find that firms with higher cyber risk earn higher average stock returns. When these firms underperform, cybersecurity experts tend to have higher concerns about cyber risk, and cybersecurity exchange-traded funds outperform. Further tests strengthen the identification of the cyber risk premium.

ACKNOWLEDGEMENTS

I want to acknowledge the generous support I received while studying here at the Broad School of Business finance department. I thank my committee members, Naveen Khanna, Hao Jiang, Andrei Simonov, and Ryan Israelsen, for their help and support during all my years here. I thank Naveen Khanna and Hao Jiang for their limitless patience and guidance. I thank William Grieser and Morad Zekhnini for their helpful comments and suggestions. Finally, I thank my wife, Ting Ting, for her firm support and understanding. Without her, none of this would be possible.

TABLE OF CONTENTS

CHAPTER 1

ANTS THAT MOVE THE LOG: CRASHES, DISTORTED BELIEFS, AND SOCIAL TRANSMISSION

1.1 Introduction

A long-standing point of inquiry in asset pricing and market micro-structure research concerns the role of retail traders. On the one hand, retail traders may generate noise that provides liquidity and incentivizes informed trading, both necessary elements for financial markets to function efficiently (Grossman and Stiglitz, 1980; Kyle, 1985; Black, 1986; Barber and Odean, 2000). On the other hand, correlated sentiment among retail traders can induce modest transitory price impacts that generate limits to arbitrage (Shleifer and Vishny, 1997; Barber et al., 2008, 2009). A few key features of financial markets have likely driven the historical modesty of retail traders' price impact. Specifically, transaction costs have restricted retail trading to a small portion of market volume. Moreover, correlated sentiment among retail traders was mainly confined to herd behavior or everyday exposure to salient events along with inefficient Bayesian updating (e.g., Banerjee, 1992; Bikhchandani et al., 1998; Barber et al., 2021) rather than deliberate coordination.

While these features previously characterized financial markets, recent innovations have dramatically changed the environment for retail traders. For example, Robinhood's advent of commission-free trading in 2015, followed by major online trading platforms such as Charles Schwab, TD Ameritrade, and E-trade in 2019, relaxed retail trading costs considerably. These events partially explain the exponential growth in retail trading, now responsible for as much as 25% of stock market volume (McCrank, 2021). In addition, social media platforms such as Reddit facilitate direct coordination among retail traders. These evolving characteristics are all well represented in the "GameStop" event in 2021, whereby retail traders joined forces to drive up GameStop's stock price by 3,000% to engineer a short squeeze. The GameStop event raises two critical questions. First, has improved coordination introduced the possibility that entertainment motivates many retail traders rather than profit? Second, is the GameStop event an anomaly, or have these features allowed retail traders to become "the ants that move the log," thus potentially altering their role in financial markets?

Gaining an adequate understanding of these questions will likely require considerable theoretical and empirical analysis and, therefore, is well beyond the scope of a single study. Thus, this paper aims to provide an initial systematic exploration of this topic by employing various novel empirical techniques in various settings with granular data on Robinhood trading activity and interactions among retail traders on Reddit. First, I use the standard logit regression to estimate ex-ante crash probabilities, where a "crash" is defined as the log monthly return lower than -20%.[1] Estimating crash risk by a return threshold is informative. According to Beason and Schreindorfer (2022), 80% of the average equity premium is attributable to monthly returns below -10%. However, crashes defined as over -20% monthly return drop constitutes only 5% of all stock returns in the CRSP universe from 1996-2021. Thus, predicting ex-ante crash risk is challenging because of the relatively low frequency of crashes, making it hard to construct valid counterfactuals. I employ a novel machine-learning technique that substantially improves the predictive power of low-probability binary outcomes.

Consistently with prior literature (e.g., Jang and Kang, 2019; Atilgan et al., 2020), ex-ante crash risk is negatively correlated with future stock returns. Specifically, a one-standard-deviation increase in crash risk is associated with an approximately 50 bps drop in monthly risk-adjusted returns. The return predictability remains strong conditioning on other tail risk measures (e.g. VaR in Atilgan et al. (2020)). Moreover, when lagged sentiment is high, the overpricing of high crash-risk stocks is more severe. These results are consistent with the predictions in Brunnermeier et al. (2007), where investors underestimate the left-tail probabilities when sentiment is high and thus buy more than the rational amount. Furthermore, consistent with the theory, I document that Robinhood traders disproportionately buy stocks with high ex-ante crash risk. In contrast, institutional investors tend to sell high crash-risk stocks.

It is hard to determine the direction of causality, which perhaps even cuts both ways. That is, are retail traders merely attracted to high-tail-risk stocks? Or are they part of what creates the tail risk?

[1] The -20% cutoff is motivated by prior literature (e.g., Jang and Kang, 2019), and I explore alternative return thresholds in Appendix.

To partially unpack the potential for the latter channel, building on the recent advancement in social transmission theory (Han et al., 2022), I exploit the history of the social media platform Reddit and the first-time appearances of stock tickers on "Wallstreetbets" as a quasi-natural experiment. Specifically, I use a stacked "difference-in-differences" approach (Gormley and Matsa, 2011; Cengiz et al., 2019) to document a causal effect of investors' online conversations on the ex-ante crash risk of stocks. I partially alleviate the possible endogeneity concerns by carefully constructing a match sample and conditioning on a set of characteristics that draw retail attention. The results show that on average the crash risk of stocks increases by approximately 10% within the first three months of appearance on "Wallstreetbets".

Recent work on social transmission (Hu et al., 2021) shows that the online conversations of retail investors on "Wallstreetbets" contain information that possibly drives future stock prices on a daily basis. To bolster the previous results, I build on this work and construct a novel and plausible instrument for investment-related conversations by utilizing the entire history of Reddit posts. Through an instrumental variable estimation approach, I show that a one-standard-deviation increase in online discussions in "Wallstreetbets" is associated with an approximately 2.3% increase in ex-ante crash risk at a daily frequency, where I follow prior literature (e.g., Bollen and Whaley, 2004; Van Buskirk, 2011; Kim and Zhang, 2014; Kim et al., 2016) and use the option implied volatility $SKEW$ as the proxy for crash risk. These results corroborate the previous "difference-in-differences" framework and suggest that retail investors could cause extreme stock returns via efficient herding.

Have retail traders become the ants that move the log? This paper presents a preliminary analysis to address whether we've reached a paradigm shift in the role of retail traders. There are several unique contributions. First, to the best of my knowledge, this is the first study that conducts causal inference on retail influence on crash risk or left-tail risk. Moreover, this paper proposes a new ex-ante crash risk measure via novel methodologies.

The rest of the paper is organized as follows. Section 1.2 briefly reviews the existing literature. Section 2.2 explains the construction of ex-ante crash risk and corresponding results for estimating monthly crash probabilities. Section 2.4 conducts asset pricing tests for crash risk in the cross-section of stock returns. Section 1.5 discusses the distorted belief mechanism for the negative price of crash risk. Section 1.6 documents the causal effect of retail conversations on firm crash risk. Section 1.7 constructs a novel instrument to provide further evidence on the causal effect of social transmission on crash risk. Section 1.8 concludes.

1.2 Literature Review

This study is related to an extensive list of areas in literature. First and foremost, it concerns the firm-level crash risk. The corporate finance literature studies the determinants of firm crash risk. These determinants are often motivated by managers hoarding bad news (Jin and Myers, 2006). The idea is that the hoarding delays the information transmission such that when it is ultimately released, there is a sudden drop in the price corresponding to the size of the cumulative bad news. Motivated by this theory, the literature has proposed a list of determinants that could endogenously influence crash risk, such as earnings management (Hutton et al., 2009), tax avoidance (Kim et al., 2011), annual report readability (Li, 2008), CSR (Kim et al., 2014), liquidity (Chang et al., 2016), short interest (Callen and Fang, 2015), and governance (Andreou et al., 2016; An and Zhang, 2013). This paper differs from this literature in that it estimates crash risk at a monthly frequency, by utilizing a rich set of conditional information (Chen and Zimmermann, 2021).

In asset pricing, a rich body of literature extracts information from option prices to determine the size of tail risk. For example, Pan (2002) provides theoretical support for the jump-risk premia implied by near-the-money short-dated options that help explain volatility smirk. Xing et al. (2010) studies the relationship between implied volatility smirks and the cross-section of stock returns. They show that the difference between the implied volatility of out-of-money put options and at-the-money call options shows strong predicting power for future stock returns. Yan (2011) show that jump size proxied by the slope of volatility smile predicts the cross-section of stock returns. The present study uses option information as one set of variables in predicting crashes, thus exploiting a far richer information set.

The third strand of literature on crash risk directly predicts the probability of crashes. Chen et al. (2001) employs cross-sectional regressions to forecast the skewness of daily stock returns. Campbell et al. (2008) use a dynamic logit model to predict distress probabilities for the cross-section of firms. Conrad et al. (2014) show that high distress risk stocks are also likely to become jackpots. They use a logit model to predict the probability of deaths and jackpots. Jang and Kang (2019) exploits a multinomial logit model to jointly predict probabilities of crashes and jackpots at an annual horizon.

This study is also related to the literature on the relationship between investor trading and market efficiency and bubble formation. De Long et al. (1990a), De Long et al. (1990b), and Abreu and Brunnermeier (2003) provide the theoretical support to and empirical evidence of positive feedback traders and their potential impact on market. Retail investors are believed to be "noise traders" that trade too much (Barber and Odean, 2000). Speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar, 2013). Recent evidence from "Robinhood Traders" shows that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber et al., 2021), while there is also evidence that mimicking portfolios based on the characteristics of "Robinhood Traders" do not seem to underperform the market, but instead could be a market stabilizing force (Welch, 2020). On the pricing impact of retail trading, Foucault et al. (2011) was one of the first papers that use a quasi-natural experiment to identify the causal effect of retail trading on stock volatility.

Finally, this study is related to the emerging literature that studies the implications and applications of machine learning methodologies in asset pricing. They are mostly concerned with resolving the "factor zoo" problem (Kozak et al., 2020; Feng et al., 2020; Bianchi et al., 2021; Gu et al., 2020).

1.3 Data and Estimation of Crash Risk

I use two sets of measures for ex-ante crash risk, one monthly measure, and one daily measure. The monthly measure is the ex-ante probability of stock crashing in a certain month, while the daily measure $SKEW$ is motivated by Xing et al. (2010), and defined as the difference between the implied volatility of out-of-money put option and that of the at-the-money call option.[2] I will start by describing the monthly measure and defer the discussion of the daily measure to Section 1.7.

1.3.1 Estimation of Monthly Ex-Ante Crash Risk

I define firm-level crashes as stock monthly log returns lower than -20%. The choice is reasonable in the following sense. Prior literature uses log annual returns of -70% as the cutoff points (Conrad et al., 2014; Jang and Kang, 2019). The unconditional probabilities of crashes defined this way at the annual frequency are roughly 5%. At a monthly frequency, a cutoff point at -20% agrees with this distribution. Thus the universe of stock returns falls into two categories – crashes and otherwise. Then the monthly ex-ante crash risk is defined as follows:

$$CrashRisk_{i,t} = E[P(r_{i,t} < -20\%)|X_{i,t-j}] \tag{1.1}$$

Where r is the monthly log return. $j \in [1, 2, 3, 4, 5, 6]$ is the months in each training window, or in other words, the period we draw conditional information. X is a set of firm-level predictors.

Estimating the ex-ante probabilities of future crashes naturally calls for a logistic regression, where the dependent variable is a binary response D_{crash}, where it equals one if the log monthly return is lower than -20%, and zero otherwise. A critical issue arises, however, in forecasting rare events such as crashes. The usual logistic estimator could produce suboptimal results due to the poor finite sample properties (King and Zeng, 2001). I provide a simple intuition for this argument in Appendix **??**. Though the difficulties and the associated statistical issues in forecasting rare events are rarely studied in economics, the remedy is readily available in machine learning literature. I follow Jiang et al. (2020) and introduce an Ensemble method, "Easy Ensemble" (EEC), that combines random undersampling and bootstrapping (Liu et al., 2008) to supplement the logistic regression approach. A detailed discussion of this technique can be found in Appendix **??**.

To estimate the ex-ante probabilities of a crash, it is essential to conduct out-of-sample procedures. Thus I use a rolling window of 6 months to estimate parameters and fit the following month to produce an OOS estimate of crash risk. With respect to the independent variables, in a slight

[2] The $SKEW$ measure by Xing et al. (2010) is widely used in the corporate finance literature as a proxy for firm crash risk. See for example...

departure from prior literature, I choose a large set of characteristics that have been shown as return predictors as the independent variables in the estimation process. Specifically, I use variables obtained from Chen and Zimmermann (2021). These are monthly firm-level characteristics that have been shown in the literature as important drivers of future returns, and these variables encompass all variables that were considered as predictors of crashes (Campbell et al., 2008; Conrad et al., 2014; Jang and Kang, 2019).

I limit the data scope to between 1996 and 2020, both to reduce the computation load and to ensure maximum data usage, as some variables are only available from 1996 (for example, option variables). Therefore, with 6-month rolling windows for training, our out-of-sample prediction starts from July 1996 to December 2020, comprised of 294 months. I use CRSP for monthly stock returns. I require common stocks with a share code of 10 or 11 and with prior month-end stock prices greater than \$5 to avoid extreme outliers.

Next, I compare the usual logistic estimator with the EasyEnsemble method in forecasting performances. To illustrate the performance difference, I conduct the following experiment. For the whole sample, I plot the percentages of real crashes predicted by either model against a decision threshold from zero to one, meaning that at each threshold, all stocks with a predicted probability higher than that would be labeled "crash". The results are shown in Figure 1.1.

Note that EasyEnsemble outperforms logistic regression in the low threshold region. This result is desirable because we know that crashes are low-probability events (the unconditional probability of a crash is around 5-6%), and we want the classifier to do well in this region. For example, at the 7% threshold, meaning that we predict all stocks with a probability estimate greater than 7% to crash in the next month, logistic regression is able to capture 72% of all real crashes, while EasyEnsemble is able to capture 85%.

1.3.2 Summary Statistics

Given the refined estimate of monthly ex-ante crash risk, we can examine its relationship with firm characteristics. In particular, we are interested in the relationship between the risk and the underlying regressors. We summarize the relationship between the machine learning-

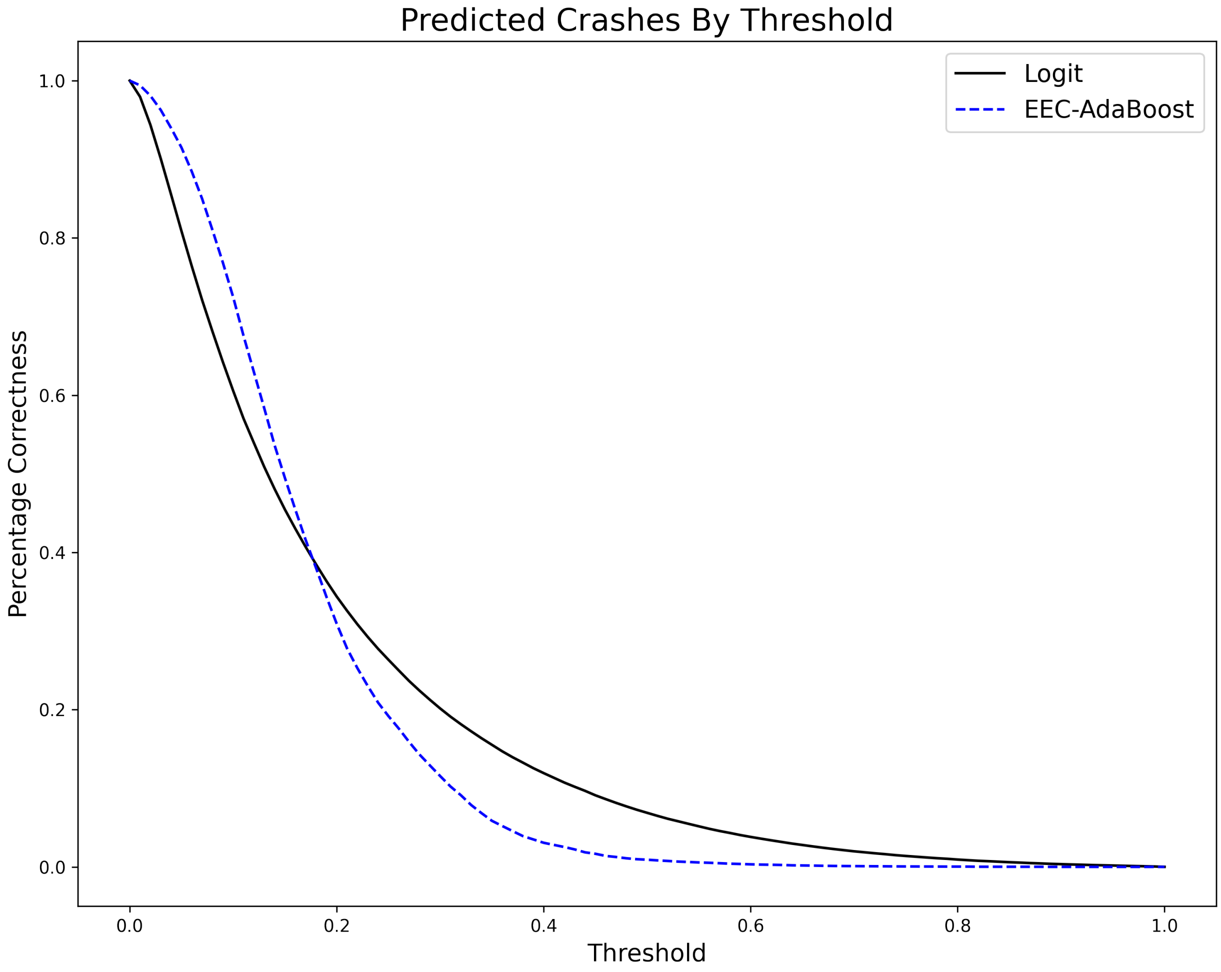


Figure 1.1 Out-of-Sample Predicted Crashes by Thresholds

The figure depicts the total percentage of out-of-sample predicted crashes for logistic regression and EasyEnsemble against decision thresholds. The X-axis is the decision threshold from zero to one. The Y-axis gives the percentage of real crashes successfully predicted by either model based on the decision threshold. For example, at the 7% threshold, meaning that we predict all stocks with a probability greater than 7% to crash in the next month, logistic regression is able to catch 72% of all real crashes, while EasyEnsemble is able to catch 85%.

generated crash risk and the top regressors in Appendix. The summary statistics of both logit-generated ex-ante crash risk and machine learning-generated crash risk, along with all relevant stock characteristics and other data used in later analyses, are presented in Table 2.1.

Table 1.1 Summary Statistics

	Crash1	Crash2	VaR1%	VaR5%	Size	Beta	Log(B/M)	ATG	GP	MOM
count	1,383,264	1,383,264	1,393,933	1,393,933	1,439,823	1,284,824	1,273,512	1,235,657	1,068,246	1,346,720
mean	0.09	0.10	-0.08	-0.05	5.73	1.08	-0.77	0.20	0.32	0.14
std	0.13	0.09	0.05	0.03	2.17	0.85	1.04	3.45	0.39	0.84
1%	0.00	0.00	-0.26	-0.16	1.33	-0.21	-3.75	-0.54	-0.78	-0.86
25%	0.02	0.03	-0.11	-0.07	4.14	0.50	-1.32	-0.03	0.15	-0.23
50%	0.04	0.07	-0.07	-0.04	5.62	0.93	-0.69	0.06	0.30	0.04
75%	0.11	0.14	-0.05	-0.03	7.18	1.48	-0.14	0.19	0.48	0.32
99%	0.64	0.42	0.00	0.00	11.08	3.72	1.72	2.64	1.25	2.89
	ST-Rev	Vol	Skew	TailBeta	Coskew	IdioRisk	Illiq	MaxRet	IO	UserNum
count	1,469,593	1,466,228	1,437,111	951,654	1,356,663	1,435,097	1,345,881	1,440,263	496,204	87,456
mean	0.01	0.03	0.24	0.72	0.22	0.03	4.72	0.08	0.41	3418.46
std	0.20	0.03	1.00	0.58	0.29	0.03	46.24	0.11	0.34	21496.13
1%	-0.43	0.00	-2.64	-0.49	-0.41	0.00	0.00	0.01	0.00	6.00
25%	-0.07	0.02	-0.28	0.37	0.04	0.01	0.00	0.03	0.09	95.00
50%	0.00	0.03	0.20	0.63	0.20	0.02	0.03	0.06	0.36	319.00
75%	0.07	0.04	0.72	0.99	0.37	0.04	0.55	0.10	0.71	1161.00
99%	0.62	0.15	3.25	2.52	1.09	0.15	86.39	0.45	1.09	64691.10

This table reports the summary statistics of our main variable ex-ante monthly crash risk and other firm characteristics used later in our analyses. There are two sets of crash risk estimates. $Crash1$ is estimated by logit regression, and $Crash2$ by machine learning (EEC-Adaboost). To differentiate our measure from the left-tail measure VaR in Atilgan et al. (2020), we also include their measure. $VaR1\%$ is defined as the 1 percentile daily return of the stock in the past year, while $VaR5\%$ is the 5 percentile daily return of the stock in the past year. Other variables include the natural log of market capitalizations ($Size$), the natural log of book-to-market ratio, asset growth (ATG), gross profitability (GP), momentum (prior 11-to-1 month returns, MOM), and short-term reversal (prior 1-month returns, $ST - Rev$), idiosyncratic volatility, illiquidity (Amihud, 2002), market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), MAX (Bali et al., 2011). IO is the institutional ownership for each stock, measured at the quarterly frequency. $UserNum$ is the total number of users for each stock on Robinhood by the end of each month. The sample starts from July 1996 to December 2021, except for Robinhood user numbers where it is limited to between May 2018 and August 2020 due to the data availability of Robintrack (https://robintrack.net/).

On top of firm-level crashes, the aggregate probability of a market crash is of great interest to researchers and practitioners alike. Although one can argue that the aggregate stock market crash is systematic, while firm-level crashes are more idiosyncratic in nature, aggregating firm-level crash probabilities might still contain information about the aggregate crash risk. One possible reason for this logic is that we use a fixed threshold (-20% log return) to define crashes, and thus aggregating these firm-level probabilities contains a systematic component. Therefore, I aggregate monthly firm-level crash risk to the market level by their lagged market capitalizations and plot the series in Figure 1.2.

On top of the aggregate crash risk series, I also plot NBER recession periods (NBER, 2021) in the gray shaded areas. Though not immediately clear, the series does contain some information about future possibilities of market crashes, as there are signs of spikes ahead of or during recession periods. Next, we move on to examine the pricing implications of firm-level monthly crash risk.

1.4 Monthly Crash Risk and Stock Returns

In this section, I examine whether the ex-ante monthly crash risk is priced in the market. I conduct both time-series portfolio analysis and cross-sectional analysis. Prior literature (Conrad et al., 2014; Jang and Kang, 2019; Atilgan et al., 2020) has indicated that crash risk, or left-tail risk, is negatively priced in the market. Though my measure is different in its time frequency and construction, we should expect similar behavior.

1.4.1 Portfolio Analysis

At the end of each month, I sort stocks into ten decile portfolios based on their estimated ex-ante crash probabilities. Then I compute both value-weighted and equal-weighted excess returns of each portfolio and the hedge portfolio that long high crash risk decile portfolio and short low crash risk decile portfolio. I regress the time series of returns on various asset pricing factors and compute the alpha estimates and their associated T-statistics. The asset pricing models include: CAPM, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and then augmented with momentum factor (FF6). To show the consistency of the results and

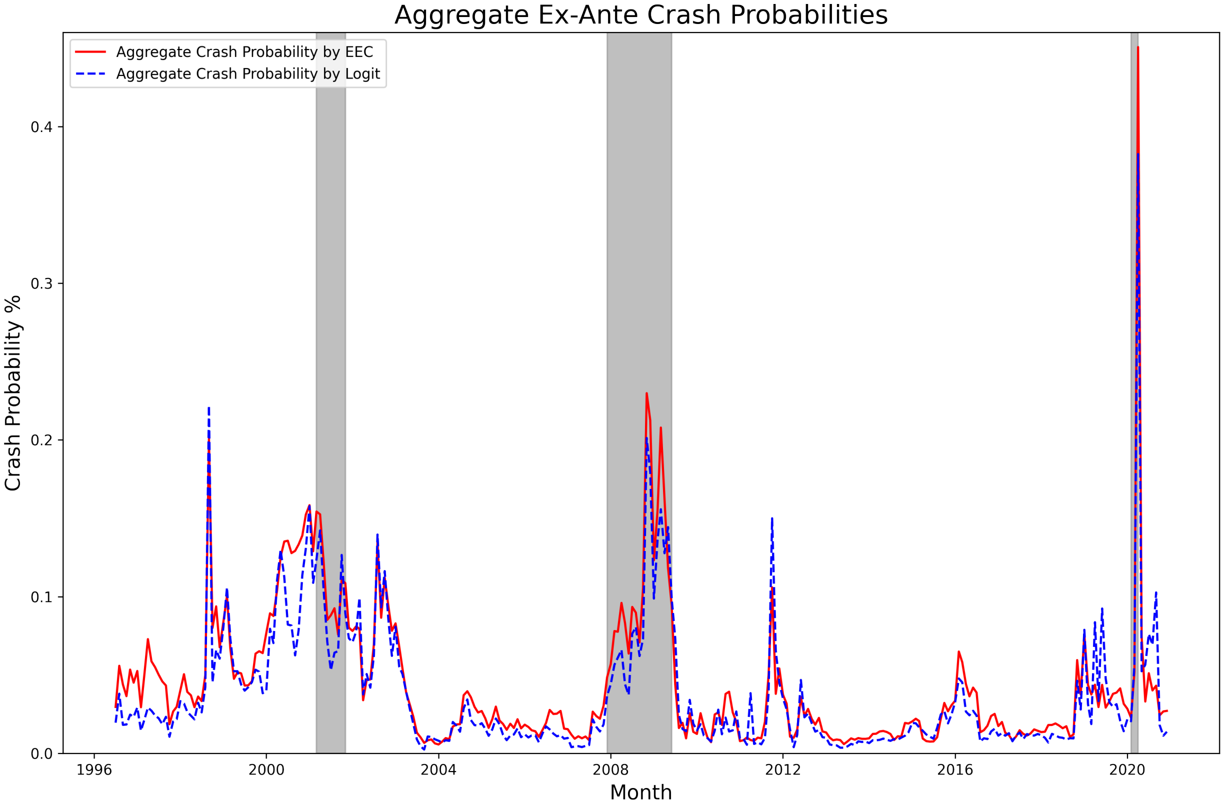


Figure 1.2 Aggregate Crash Risk

The figure plots market-wide aggregate ex-ante crash probabilities from 1996 to 2020. The aggregation is done by weighting the monthly ex-ante crash risk of each firm by their lagged market capitalizations as follows:

$$AggCrashRisk_t = \frac{\sum_i MarketCap_{i,t-1} \times CrashRisk_{i,t}}{\sum_i MarketCap_{i,t-1}}$$

The red solid line indicates the aggregate crash probabilities by using the machine learning-generated crash probabilities, while the blue dashed line uses the logit-generated crash probabilities. The gray shaded areas indicate NBER recession periods (NBER, 2021). The time series run from July 1996 to December 2020.

the superiority of the EasyEnsemble method, I show alpha estimates using both logistic regression and EasyEnsemble in Table 1.2.

Table 1.2 Decile High-Minus-Low Portfolio Alphas

		Logit		EEC-Adaboost	
	Pricing model	Alpha	T-stat	Alpha	T-stat
Value-weighted	CAPM	-1.852	-3.730	-1.967	-4.393
	FF3	-1.842	-4.440	-1.963	-5.456
	FF4	-1.533	-3.531	-1.775	-4.636
	FF5	-0.874	-2.834	-1.120	-3.947
	FF6	-0.696	-2.263	-1.023	-3.442
Equal-weighted	CAPM	-2.470	-5.571	-2.458	-5.325
	FF3	-2.461	-7.941	-2.452	-7.573
	FF4	-2.106	-7.161	-2.173	-7.005
	FF5	-1.656	-5.637	-1.783	-6.093
	FF6	-1.438	-5.788	-1.614	-5.947

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

This table presents the analysis of portfolios sorted on the ex-ante crash risk measures estimated by both logit and machine learning (EEC-AdaBoost). At the end of each month, stocks are ranked by their ex-ante crash probabilities produced by either logit or machine learning into ten decile portfolios. Then we compute both equal-weighted portfolio returns and value-weighted returns by their lagged market capitalization. The hedge portfolio is long in the top decile ex-ante crash risk portfolio and short in the bottom decile crash risk portfolio. Then the hedge portfolio return series are regressed on risk factor returns from various empirical asset pricing models. The asset pricing models include: CAPM, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and then augmented with momentum factor (FF6). Then we report the resulting intercepts (alphas) and their associated T-statistics. The upper panel presents results from using value-weighted portfolio returns, while the lower panel presents equal-weighted results. The left half shows results from using ex-ante crash risk estimated from logistic regressions, and the right from machine learning (EEC-AdaBoost). Standard errors are adjusted using the Newey-West procedure (Newey and West, 1986) with 6 lags.

As shown in Table 1.2, when we long top crash risk decile portfolio and short bottom decile portfolio, we produce consistent and significant negative alphas across different asset pricing models, equal-weighted or value-weighted, with T-statistics of magnitude well over 3. Note also that when we compare the results from using logit-generated crash risk and machine learning-generated crash risk, the latter shows superiority in both the magnitude of alpha and the T-statistics. This is a strong piece of evidence that machine learning not only produces consistent results with conventional methods but also demonstrates better forecasting efficacy, as it classifies correctly more actual crashes that contribute to lower returns in the subsequent month.

1.4.2 Cross-Sectional Regressions

Next, I run Fama-MacBeth cross-sectional regressions (Fama and MacBeth, 1973a) following the procedure in Fama and French (2020). Each month, I regress raw stock returns on cross-sectionally standardized lagged firm characteristics. Then I average the coefficients to arrive at the final estimates. The coefficients on characteristics can be directly interpreted as average priced return spread for one standard deviation increase of the corresponding firm risk. I include common risk characteristics such as the natural log of market capitalizations, natural log of book-to-market ratio, asset growth, gross profitability, momentum (prior 11-to-1 month returns), short-term reversal (prior 1-month returns), and my estimated crash probabilities from the Ensemble method. On top of these variables, I control for a set of anomaly characteristics that are shown to be significantly correlated with future stock returns: idiosyncratic volatility, illiquidity (Amihud, 2002), market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), and net operating assets NOA (Hirshleifer et al., 2004). Bali et al. (2011) proposes a measure MAX that represents investors' preference for lottery-like payoffs. MAX stands for the maximum daily return achieved by each stock in the prior month. To see if the estimated crash risk carries additional information that distinguishes it from MAX, I add the MAX measure as a control variable in the Fama-MacBeth regressions.

Atilgan et al. (2020) also studies the left-tail risk, although their measure is constructed differently. Their "value-at-risk" (VaR) is entirely based on historical returns and is defined as the return conditioning on probability distribution, which differs from our measure that takes return cutoff as given and estimates ex-ante probabilities. To see whether our crash risk contains incremental information about future stock returns than the VaR measure, I include VaR as a control variable. The VaR measure is the negative of 1 percentile daily return of the stock in the past year. I report the regression results in Table 1.3.

Table 1.3 Fama-MacBeth Cross-Sectional Regressions

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Returns in %				
Crash Risk (Logit)	-0.491***	-0.453***			
	(0.080)	(0.077)			
Crash Risk (EEC)			-0.507***	-0.459***	
			(0.097)	(0.086)	
VaR1%		-0.123		-0.097	-0.246***
		(0.082)		(0.074)	(0.083)
Controls	YES	YES	YES	YES	YES
Observations	545,367	545,290	545,367	545,290	564,466
R-squared	0.083	0.086	0.083	0.085	0.084
Note:					*p<0.1; **p<0.05; ***p<0.01

This table reports Fama-MacBeth cross-sectional regressions of raw returns on ex-ante crash risk and lagged firm characteristics in the spirit of Fama and French (2020). First, we regress monthly stock returns of each month on lagged firm characteristics. Then we average the coefficients and report the associated standard errors. Our main variables of interest are the two ex-ante crash risk measures. One is estimated by logit regression, and the other by machine learning (EEC-Adaboost). Columns (1) and (2) use the logit-generated crash risk as the main variable, while Columns (3) and (4) use the machine learning-generated crash risk. To differentiate our measure from the left-tail measure VaR in Atilgan et al. (2020), I include their measure in Columns (2) and (4) as a control variable, where $VaR1\%$ is defined as the negative of 1 percentile daily return of the stock in the past year. In Column (5), I only include $VaR1\%$ as the sole variable to proxy for left-tail risk to ensure that our results are consistent with Atilgan et al. (2020). Other control variables include the natural log of market capitalizations, the natural log of book-to-market ratio, asset growth, gross profitability, momentum (prior 11-to-1 month returns), and short-term reversal (prior 1-month returns). In Column (3), I add idiosyncratic volatility and illiquidity (Amihud, 2002). In Column (4), I add market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), net operating assets NOA (Hirshleifer et al., 2004), and MAX (Bali et al., 2011). All independent variables are standardized cross-sectionally each month to be mean zero and standard deviation of unity, such that the coefficients on all the independent variables can be directly read as the percentage increase in average stock returns if the underlying independent variable increase by one standard deviation. Standard errors are adjusted according to Newey-West procedures (Newey and West, 1986) with 6 lags.

Table 1.3 suggests several points. First, both logit-generated ex-ante crash risk and machine learning-generated crash risk are significantly and negatively correlated with future stock returns, and their magnitudes are very similar to each other. Second, the loadings on crash risk are robust even after controlling for common risk characteristics and go beyond a plethora of tail risk-related variables, including the lottery-payoff proxy MAX (Bali et al., 2011). Third, when our crash risk is not included in the regression, the VaR measure is significantly and negatively correlated with future stock returns, consistent with the results in Atilgan et al. (2020). However, when our crash risk is included in the regression, the loading on VaR becomes insignificant, while our crash risk measure loads negative and significant consistently. This suggests that both our logit-generated and machine learning-generated crash risk measures contain more information than VaR and consequently subsume its effect. Depending on the control variables and the measure we use, a one-standard-deviation increase in ex-ante monthly crash risk is associated with approximately a 45-51 bps drop in subsequent risk-adjust returns, which translates into -5.47% to -6.12% in annual risk-adjusted returns. These results corroborate the prior literature that ex-ante crash risk is negatively priced, and also provide strong evidence that our crash risk measure contains richer and incremental information than existing crash risk measures.

1.5 A Possible Economic Mechanism: Distorted Belief

The negative price of crash risk does not agree with rational expectations, as a rational investor would naturally demand a positive risk premium for holding such risk. Prior literature attempts to explain the phenomenon via several arguments. One argument is the limits to arbitrage (Shleifer and Vishny, 1997; Conrad et al., 2014; Jang and Kang, 2019). They show evidence that institutional investors tend to "ride the bubble" as rational speculators, instead of trading against crash risk as rational arbitragers, since high crash risk stocks tend to be small, illiquid, and hence costly to short. The second argument is that investors underestimate the momentum in the left tail (Atilgan et al., 2020), meaning that stocks that crashed the last month may well be highly possible to continue crashing in the subsequent month. Investors somehow fail to understand this dynamic and "bought the dip", which renders the stocks with high crash probabilities overpriced. However, it is unclear why this momentum exists. Moreover, since the VaR measured used Atilgan et al. (2020) is an ex-post measure, it does not answer the question from an investor behavior perspective. A third argument pertains to the observation that stocks with extreme past returns are attention-grabbing, and retail investors have a preference for such stocks (Barber and Odean, 2008; Barber et al., 2021). However, it is reasonable to assume that investors are drawn to extreme past winners, as they might be over-extrapolating past returns. It is nonetheless puzzling why investors should prefer extreme past losers. Moreover, it is difficult to understand why investors should prefer high left-tail stocks. Even if they underestimate the momentum in the left tail, these are undesirable stocks from a risk-return tradeoff standpoint. In addition, over time, investors should be able to learn from past observations that high crash risk stocks are overpriced, as many of them indeed crashed in the subsequent month.

The literature in behavioral theories provides valuable guidance in terms of investor beliefs and preferences towards crash risk or left tail risk. Two theories, in particular, have clear predictions about investors' attitudes towards the left tail. One is cumulative prospect theory (CPT) by Barberis and Huang (2008). They show that investors with a CPT preference would overweight small probability events. One example is that people would gamble on slim chances of big payoffs, but buy insurance for plane crashes. The implication is that investors with CPT preference should shun high crash risk stocks since they effectively deem those crashes more likely to happen than the true distribution. If all investors have such a preference, high crash risk stocks should be underpriced, and thus produce a positive risk-adjusted return. This prediction does not seem to conform to the empirical observation.

The second theory is the optimal expectations theory (OET) by Brunnermeier et al. (2007). They show that investors may derive anticipatory utility when holding an optimistic subjective belief about stock returns, even though such beliefs prove to be wrong afterward. If investors hold such a belief, they would effectively shift their subjective return distribution to the right when their sentiment is high. The implication is that when sentiment is high, investors with such beliefs tend to think that crashes are less likely than reality, and thus overbuy high crash risk stocks. The pricing implication is that crash risk or left tail risk is overpriced and thus predicts a negative risk-adjusted return.

The evidence presented in this paper and the prior literature for the negative price of crash risk agrees with the optimal expectations theory. To further establish evidence as to whether investors overbuy high crash risk stocks when their sentiment is high, I conducted several tests to provide additional evidence.

1.5.1 Crash Risk Portfolio Returns and Sentiment

First, I examine the relationship between the crash risk hedge portfolio returns and sentiment. If investors hold optimal expectations, then the loss on the crash risk high-minus-low hedge portfolio would be higher when lagged sentiment is high, since investors' belief distortion would be more severe during such periods.

I follow Baker and Wurgler (2006) and use their sentiment index as a proxy for the market-wide sentiment. In particular, I use the sentiment measure that is orthogonal to macroeconomic indicators to alleviate the impact of market risks. Since their index is available up to the year 2018, my sample is hence limited between July 1996 and December 2018. Then I divide the sample period into two subperiods, where one is the high sentiment period when sentiment is higher than the median value of the whole sample, and another is the low sentiment period. Then I compute the excess returns of the top decile portfolio, the bottom decile portfolio, and the long-short hedge portfolio that long high crash risk stocks and short crash risk stocks, in each of the subperiods. I then compute the differences in these returns between high and low sentiment periods. The results are summarized in Panel A of Table 1.4.

It is immediately clear from the table that the high-crash-risk stocks experience the lowest returns after a high sentiment period when mispricing is most severe, while they do not show negative returns on average after low sentiment months. On the other hand, there is no statistically significant difference between high and low sentiment periods for low-crash-risk stocks. On the whole, a long-short strategy that is long high-crash-risk stocks and short low-crash-risk stocks produces more negative and significant excess returns after high sentiment months. These results are consistent with our hypothesis that when investors are bullish, they are more likely to overbuy high-crash-risk stocks, and thereby the expected returns of these stocks would be lower.

Table 1.4 Sentiment and Crash Risk Returns

Panel A: Portfolio Excess Returns and Sentiment

	High Sent	Low Sent	High-Low
Low crash risk	0.597*	0.812**	-0.215
	(0.314)	(0.309)	0.436
High crash risk	-1.849*	0.879	-2.728**
	(1.008)	(0.880)	(1.280)
Long-short	-2.446**	0.067	-2.513**
	(0.943)	(0.709)	(1.144)

Panel B: Price of Crash Risk and Sentiment

	(1)	(2)	(3)	(4)
	FMB		Panel	
VARIABLES	Low Sent	High Sent	Return	Return
Crash Risk	-0.405***	-0.619***	-0.335***	-0.135**
	(0.108)	(0.141)	(0.050)	(0.062)
SentmentD×Crash Risk				-0.374***
				(0.063)
Controls	YES	YES	YES	YES
Observations	240,805	269,577	545,227	510,260
R-squared	0.078	0.085	0.168	0.159

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

This table presents the relationship between the price of crash risk and sentiment. Panel A reports the value-weighted portfolio excess returns for high-crash-risk, low-crash-risk, and long-short hedge portfolios in both high sentiment and low sentiment periods and their differences. Market-wide sentiment is defined by Baker and Wurgler (2006). Our sample is limited between July 1996 and December 2018 because of the availability of the index. High and low sentiment periods are defined as either above or below median sentiment over the sample period. Panel B reports regression results. In Columns (1) and (2), we run Fama-MacBeth cross-sectional regressions of stock returns on crash risk and lagged firm characteristics for high- and low-lag-sentiment periods separately. Control variables include all the firm characteristics used in Table 1.3. Standard errors are estimated according to Newey-West procedure (Newey and West, 1986) with 6 lags. In Columns (3) and (4), we run panel regressions of stock returns on the same independent variables as the previous specification, with firm and time fixed effects. In Column (4), we add a dummy variable $SentD$, where it equals one if the lagged sentiment is higher than the sample median, and zero otherwise. We interact $SentD$ with crash risk, and hence the coefficient on the interaction term can be interpreted as the incremental price of crash risk when lagged sentiment is high. All independent variables are standardized cross-sectionally each month to be mean zero and standard deviation of unity. Standard errors are clustered at the firm level.

To further examine the relationship between crash risk and sentiment, I run Fama-MacBeth regressions and panel regressions of stock returns on firm characteristics for high- and low-lag-sentiment months separately. The hypothesis is that the price of crash risk should be more negative immediately after high sentiment months. As before, high sentiment months are defined as those months with lag sentiment higher than the sample median, and low sentiment months are defined as those months with lag sentiment lower than the sample median. The results are reported in the first two columns in Panel B of Table 1.4.

We can see from the table that when lagged sentiment is high, the coefficient on crash risk is -0.619%, compared to -0.405% when lagged sentiment is low. In other words, the price of crash risk associated with a one-standard-deviation increase in the risk is 21 bps lower when lagged sentiment is high. Though the difference between the two coefficients is not statistically significant (T-statistic of -1.2), the annualized return difference is large at -2.52%. This is another piece of evidence that high crash risk stocks are more overpriced when lagged sentiment is high.

To further assess this phenomenon, I also conduct the following analysis. I define a dummy variable $SentD$, where it equals one if the lagged sentiment is higher than the sample median, and zero otherwise. I first run a panel regression of stock returns on crash risk and other firm characteristics, with firm and time fixed effects. Then I include the $SentD$ variable and interact it with crash risk. The hypothesis is that the interaction term should be significantly negative since when lagged sentiment is high, the overpricing of high crash risk stocks should be more severe. I report the results in Columns (3) and (4) in Panel B of Table 1.4.

As shown in the table, even after including firm and time-fixed effects, the ex-ante crash risk is consistently priced negatively, albeit with a smaller magnitude. In Column (4), when we interact the sentiment dummy with crash risk, the loading on crash risk is much smaller in magnitude and statistically significant at 5% level, while the coefficient on the interaction term is negative and statistically significant at 1%, with a much higher magnitude. These results are consistent with our hypothesis that when lagged sentiment is high, investors buy more high crash risk stocks, which causes the overpricing of these stocks even higher, and therefore the subsequent returns turn out to be much lower than in low lagged sentiment periods.

1.5.2 Trades on Crash Risk

Next, we examine whether some investors are likely to buy high-crash-risk stocks. This hypothesis is the underlying assumption of the previous literature that high-crash-risk stocks are overpriced and is an implication from (Brunnermeier et al., 2007). To explore this hypothesis, I first use Robintrack data to construct a retail trading measure and examine whether they tend to buy high-crash-risk stocks.[3]

As has been extensively discussed in Barber et al. (2021) and Welch (2020), Robintrack data contains hourly stock popularity numbers that are measured by how many users on Robinhood hold a particular stock at a certain hour. Since we cannot observe the number of shares they hold for each stock, and there is no data for the total number of users for each time period, the next best solution is to measure the change in the number of users for each stock. As crash risk is estimated at a monthly frequency, I use month-end numbers of Robinhood users to merge the data. I first construct a log measure for Robinhood trading:

$$Change\,in\,Log(\#User_{i,t}) = \log(\#User_{i,t}) - \log(\#User_{i,t-1}) \tag{1.2}$$

Then I follow Barber et al. (2021) and construct a percentage change measure for Robinhood trading:

$$\%Change\#User_{i,t} = \#User_{i,t}/\#User_{i,t-1} - 1 \tag{1.3}$$

Where t is at the monthly frequency to match the frequency of our ex-ante crash risk measures. The specification is as follows:

$$Robinhood\,Trade_{i,t} = \alpha_0 + \beta \times Crash\,Risk_{i,t} + \sum_p \beta_p Control_{p,i,t-1} + \alpha_i + \lambda_t + \epsilon_{i,t} \tag{1.4}$$

Where we add firm and time fixed effects to account for unobserved heterogeneity that might be correlated with the error term. The Robinhood sample runs from May 2018 to August 2020. I regress the Robinhood trading measures on both measures of ex-ante crash risk, controlling for the

[3] Robintrack: https://www.robintrack.net/.

lagged log of the user number and a set of firm characteristics. The results are reported in Columns (1) to (4) of Table 1.5.

Table 1.5 Investor Trading and Crash Risk

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Change in Log(User)		User%Change		IO Change	
Crash Risk (Logit)	0.093*** (0.010)		0.154*** (0.020)		-0.026*** (0.002)	
Crash Risk (EEC)		0.104*** (0.016)		0.156*** (0.028)		-0.013*** (0.003)
Controls	YES	YES	YES	YES	YES	YES
Observations	63,692	63,692	63,692	63,692	375,339	375,339
R-squared	0.241	0.240	0.191	0.190	0.500	0.500
Firm & Time FE	YES	YES	YES	YES	YES	YES

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

This table presents results from regressing Robinhood user trading measures and institutional trading measures on crash risk, controlling for other firm characteristics. The first Robinhood user trading measure is the monthly change of the natural log of user numbers holding a particular stock, where the user numbers are from the online brokerage Robinhood (Robintrack). The second Robinhood user trading measure is the percentage change in the number of users over the previous month. The institutional trading measure is the quarterly change in the ratio of institutional holding for each stock. We regress all of these trading measures on the contemporaneous crash risk measures constructed from both logit regressions and the machine learning method (EEC-AdaBoost). Columns (1) to (4) add lagged log of the number of users as a control variable. For all specifications, the control variables include the natural log of market capitalization, the natural log of book-to-market ratio, asset growth, gross profitability, momentum, short-term reversal, MAX and MIN (Bali et al., 2011), defined as the highest and lowest daily returns of the previous month, total skewness of daily returns in the previous month, illiquidity (Amihud, 2002), and Fama-French three-factor betas estimated from the previous month. Firm and Time fixed effects are included, and robust standard errors are included in parentheses.

The table shows that over the sample period when Robinhood data is available, retail investors on average tend to buy high-crash-risk stocks, consistent with our hypothesis. Importantly, in all specifications, we control for such commonly used lottery characteristics as MAX and MIN (Bali et al., 2011), which are defined as maximum and minimum daily returns of the previous month, and total skewness of the previous month. The coefficient on crash risk is consistently and significantly positive in Robinhood trading tests, meaning that retail preference for high-crash-risk stocks goes beyond the conventional proxies for lottery characteristics defined in the literature (Barberis and Huang, 2008; Bali et al., 2011).

A related question arises as to whether institutional investors would be liquidity providers and act as counterparties since literature has shown that they are reluctant to short the left tail, and would rather ride the bubble. I examine this issue by regressing the change of institutional holdings on the same set of characteristics. The institutional holdings data comes from Thomson Reuters 13F filings data and is defined as the percentage of shares held by institutional investors. The change in the holdings is the difference between the current quarter's holdings and the previous quarter's. The results are shown in Columns (5) and (6) in Table 1.5.

The results show that there is strong evidence that institutions might be the counterparty of retail investors for crash risk. In sharp contrast to Robinhood trading results, the coefficients on both crash risk measures are negative and statistically significant for institutional trading tests. Taken together, these results support the hypothesis that retail traders derive anticipatory utilities from distorted subjective beliefs. Consistent with the predictions in Brunnermeier et al. (2007), when lagged sentiment is high, investors underestimate left-tail risks and tend to overbuy stocks with high crash risk, which in turn drives up their prices, leading to lower expected returns subsequently. Both the pricing results and retail trading results conform to this theory.

1.6 Retail Influence on Monthly Crash Risk

Evidence from the previous section shows that retail investors tend to buy high ex-ante crash risk stocks, and this effect is over and beyond the effect of the usual proxies for lottery characteristics. These buying activities could be inconsequential if retail investors are pure "noise traders" (De Long et al., 1990a), as their trades are idiosyncratic and would be canceled out on average. However, when their trades are correlated because of attention or herding, they could forecast subsequent returns (Barber and Odean, 2008; Barber et al., 2021). Social media is instrumental in facilitating herding behavior, as it transmits trading strategies more efficiently. As implied in Han et al. (2022), there is an inherent feedback loop in correlated trading and asset prices. When investors (receivers) take note of other investors' (senders) recent trading success, as demonstrated by their bragging on social media about the high recent returns of their stock picks, they continue to trade in the same direction, thus pushing the stock price even higher. The implication is that regardless of whether investors display a preference for skewness, their trading actions would produce such results and influence stock prices.

There is causal evidence that suggests higher participation by retail investors does induce higher stock volatility (Foucault et al., 2011). They may be marginal price setters for small stocks (Graham and Kumar, 2006). Retail short sellers predict negative future returns, and they seem to have superior knowledge of small firm fundamentals (Kelley and Tetlock, 2017). Much of the literature focuses on predictive tests, as it is extremely difficult to find ideal settings for the proper identification of causality.

I explore a particular shock to the retail attention and herding channel that might have influenced retail investors' trading behavior, which in turn could drive the change in the crash risk of the underlying stocks.

1.6.1 The Advent of Wallstreetbets

"Wallstreetbets" is a "Subreddit" on the social media platform "Reddit", and has garnered considerable attention from the investment community largely because of the "GameStop" saga. The Subreddit started in April 2012, and today it has over 12 million subscribers. These subscribers call themselves "degenerates", and frequently exchange trading ideas and post their gains and losses. In a recent study, Hu et al. (2021) shows that conversations on "Wallstreetbets" have information content that predicts next-day returns. A study from a different discipline, Li and Wu (2018) shows that retailers displaying past sales numbers can induce consumers to herd and buy more of the products. These studies suggest that social media as a platform for idea sharing can facilitate more efficient herding. Therefore, it is conceivable that the advent of a highly efficient platform for sharing ideas might affect asset prices, including the crash risk of the underlying stocks, following the results that retail investors exacerbate the overpricing of high-crash-risk stocks.

I examine this issue by tracing back to the origin of "Wallstreetbets" when it was founded in

April 2012. I obtain and process all posts from April 2012 when the Subreddit started till December 2020, and find out all stock tickers that were mentioned in these posts.[4] I drop all ticker names that are also common English words, slang, and abbreviations. To illustrate the growing community on “Wallstreetbets”, I plot the number of posts each month that mention ticker names, and also the number of unique ticker names mentioned each month in Figure 1.3.

Panel A of Figure 1.3 plots the number of posts that mention ticker names on the Subreddit “Wallstreetbets”, and Panel B plots the number of unique tickers/firms each month. The time series spans from April 2012 when “Wallstreetbets” was started to December 2020. It shows that the activities on “Wallstreetbets” exploded after the pandemic began in 2020. It also shows the growing breadth of retail investor interests in the number of stocks.

1.6.2 The Staggered First Appearances of Stock Tickers

Members started to mention stocks in their posts on “Wallstreetbets” on the first day of the Subreddit. According to Han et al. (2022), people are more likely to mention certain stocks if these stocks happen to have high past returns. If other people see these posts, they are more likely to follow suit and trade in the same direction. This could in turn affect the stock returns.

To test this hypothesis, I focus on the seven-month window around each “event”, where “event” means a stock ticker appeared for the first time on “Wallstreetbets”. Thus there are three months pre-event, and three months post-event. Since conversations about stocks are not exogenous per se, we need a matching strategy and control variables that can offset the endogenous portion of the test. Therefore, I use propensity score matching by running logistic regression. The response is a dummy variable $D_{i,t} = 1$ if a stock i appears on “Wallstreetbets” for the first time at time t, or zero otherwise. The independent variables include lagged market capitalization, prior-month return, asset growth, book-to-market ratio, gross profitability, idiosyncratic risk, illiquidity, MAX, and prior 12-month return, to proxy for the common stock characteristics.

The estimated parameters are then fit to the whole sample to generate fitted values as the propensity score for each stock at each point in time. To match each event, I use the score generated

[4] The complete history of Reddit comments data comes from https://files.pushshift.io/reddit/comments/.

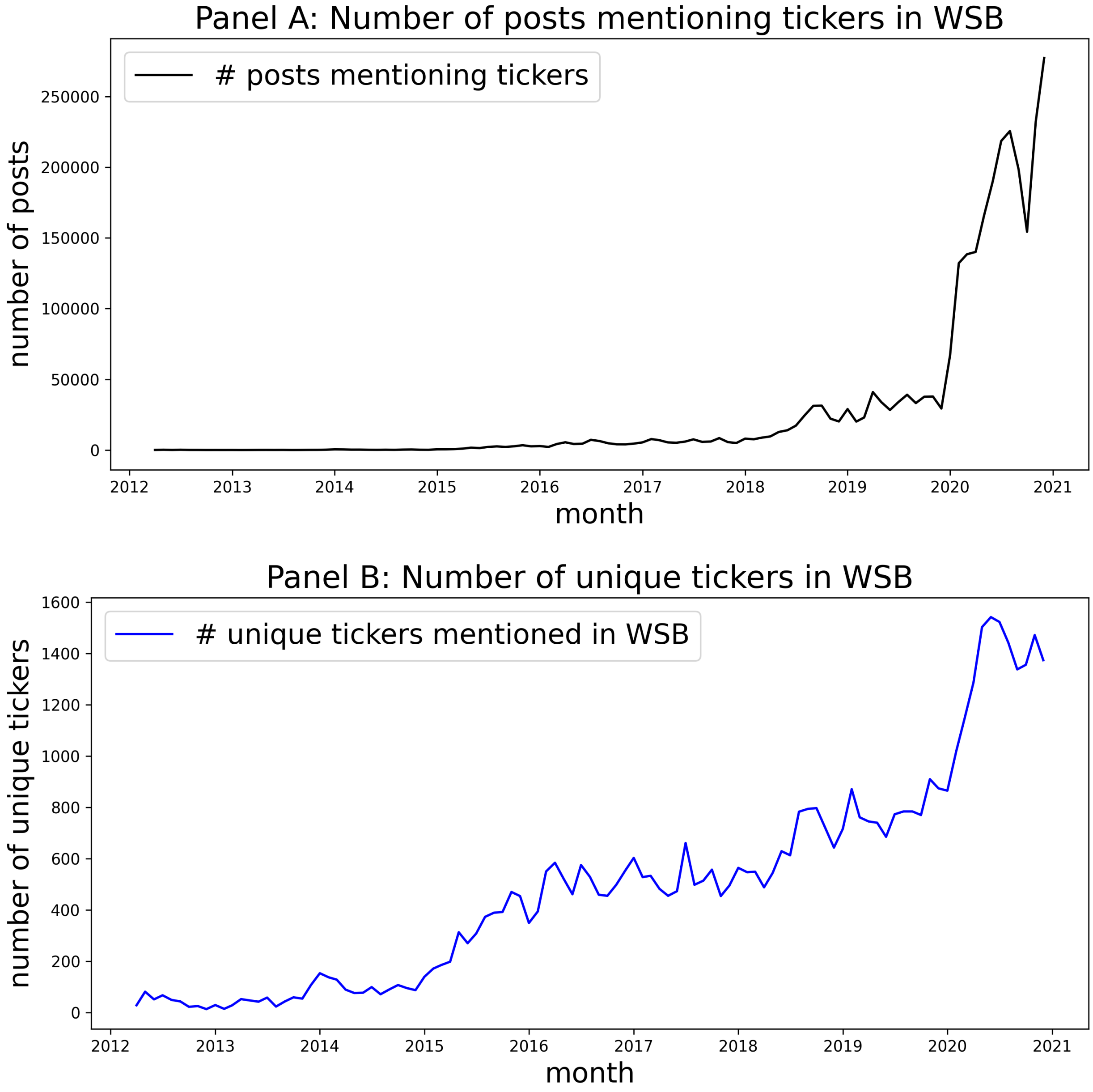


Figure 1.3 Monthly Number of Posts and Unique Tickers on “Wallstreetbets”

The figure plots the total number of posts each month on the Subreddit “Wallstreetbets” that mention stock ticker names, and also the number of unique ticker names mentioned each month in Panel A and Panel B respectively. For a ticker to be counted, it must not be common English words, slang, or abbreviations. The time series spans from April 2012 when “Wallstreetbets” was established to December 2020.

for each "never treated" stock three months prior to the event and find five stocks that have the closest propensity scores to each treated stock.[5]

After the matching process, I follow Gormley and Matsa (2011); Cengiz et al. (2019) and stack each event cohort, where each cohort contains the treated stock and the matched sample. Then I run the following specification:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_{p} \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \tag{1.5}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Column (1) and Column (3) of Table 1.6, where Column (3) adds control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Standard errors are clustered at the unit level.

When control variables are not included, there is a 1.03 percentage point estimated increase in logit-generated crash risk when a stock is first mentioned on "Wallstreetbets", and the coefficient is highly statistically significant. When control variables are included, the magnitude reduces to approximately 56 bps, and the coefficient remains statistically significant at the 1% level. This corroborates our hypothesis that when a stock was mentioned on social media and subsequently draws more attention that possibly induces more correlated retail trading, which could increase stock crash risk.

A critical assumption for the difference-in-differences analysis is the "parallel trend" assumption, where the treated group and the control group should not have significant differences before the event happens. To examine this "parallel-trend" assumption, I conduct a dynamic approach, where

[5] "Never treated" means the stock never appears on "Wallstreetbets". This is to ensure the cleanest matching. There are in total 2,276 unique stocks that are never mentioned on "Wallstreetbets".

Table 1.6 First Appearances of Stock Tickers on "Wallstreebets" and Crash Risk

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Crash Risk (Logit)			Crash Risk (EEC)		
Treated	1.032***	0.560***		0.674***	0.303***	
	(0.103)	(0.129)		(0.054)	(0.064)	
Month -3			0.009			0.001
			(0.160)			(0.082)
Month -2			-0.041			0.041
			(0.140)			(0.074)
Month 0			0.464***			0.152**
			(0.136)			(0.076)
Month +1			0.326*			0.152
			(0.185)			(0.097)
Month +2			0.689***			0.478***
			(0.199)			(0.095)
Month +3			0.735***			0.508***
			(0.218)			(0.105)
Observations	208,502	125,734	125,734	208,502	125,734	125,734
R-squared	0.874	0.909	0.909	0.921	0.946	0.946
Cohort×Units FE	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES

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

This table reports results from a "stacked difference-in-differences" approach (Gormley and Matsa, 2011) that examines the effect of first appearances of stocks tickers on "Wallstreetbets" on their ex-ante crash risk. Columns (1) to (3) use logit-generated crash risk as the dependent variable, while Columns (4) to (6) use machine learning-generated crash risk. "Wallstreetbets" was started in April 2012. From the beginning of "Wallstreetbets" to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of "never treated" stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the "cohorts" containing treated and control observations are stacked together and the following specification is run:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Columns (1), (2), (4), and (5), where Columns (2) and (5) add control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity, MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Columns (3) and (6) examine the dynamic treatment effects around the events. Standard errors are clustered at the unit level.

instead of examining the coefficient on the treatment dummy, I run the following specification:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \qquad (1.6)$$

Where the dummy variables $D_{i,j,c,t}$ indicate whether a stock i is treated in cohort c at time t, and the distance $j \in [-3, 3]$ from the current month to the treatment month. Month -1 is chosen as the base month that will be omitted from the regression. The results are included in Column (3) and (6) of Table 1.6.

As shown in the table, the coefficients for the two months before the event are economically and statistically insignificant. On the other hand, the coefficients on the treatment month and the months after the treatment are economically and statistically significant. These results provide strong support to the assumption that there are no significant differences between treatment and control groups before the treatment.

To provide further evidence of the "parallel trend" assumption, I also plotted the coefficients on the dummy variables $D_{i,j,c,t}$ with their 95% confidence intervals in Figure 1.4.

The figure provides visual support for the "parallel trend" assumption for our "difference-in-differences" analysis. The dynamic results, together with the static results, provide strong evidence that there is a possible causal effect of increased retail attention on stock crash risk.

1.6.3 Size and Institutional Ownership

Foucault et al. (2011) show that retail investors have an outsized impact on stock volatility, especially for smaller stocks, where the standard limits to arbitrage argument apply (Shleifer and Vishny, 1997). Smaller stocks are traded thinly and thus are less liquid. Because of their price tag, they are usually the preferred habitat of retail investors, and thus their institutional holding is usually lower. As a result, their prices can stay distant from their fundamentals for an extended period of time, since rational investors are reluctant to arbitrage for the arbitrage would be costly.

The same argument should apply to crash risk. Prior literature has shown that high crash risk stocks tend to be smaller and more costly to arbitrage (Jang and Kang, 2019). We have also shown in Section 1.5 that retail investors seem to display a preference for high crash risk stocks possibly

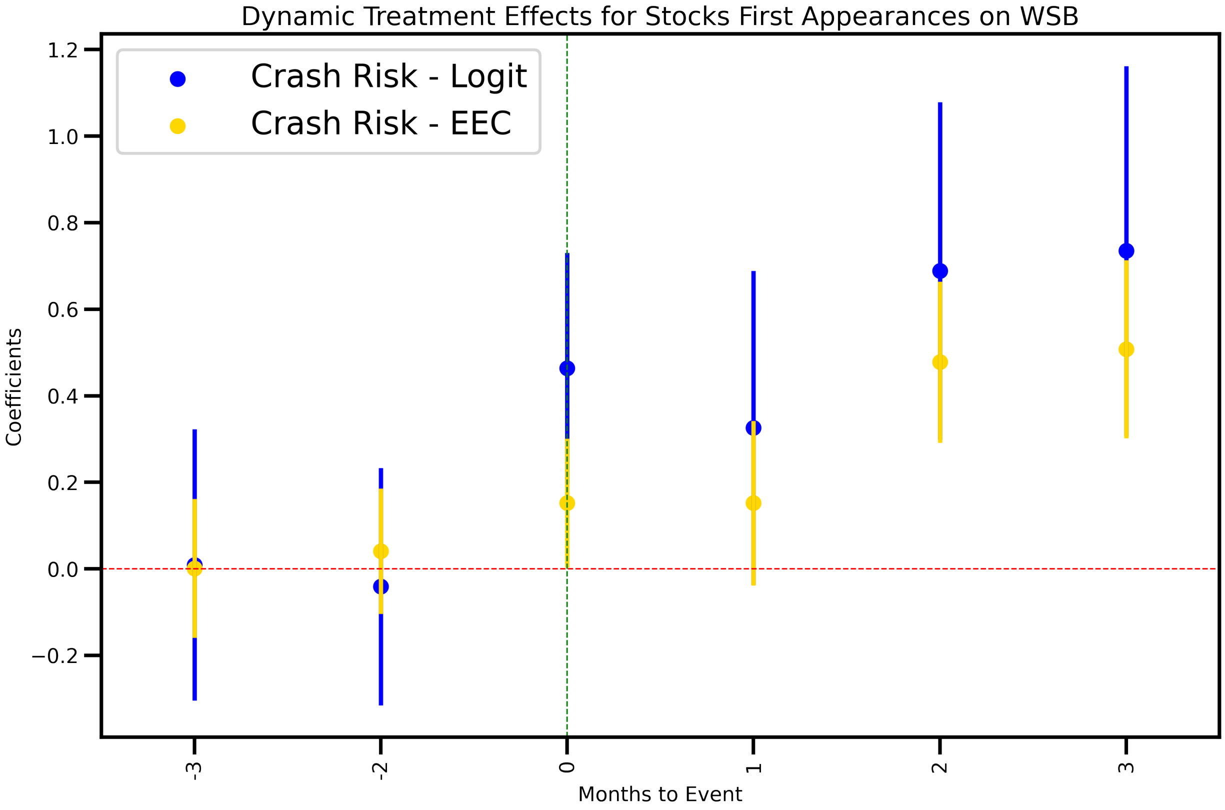


Figure 1.4 Dynamic Treatment Effects of the First Appearances of Tickers on “Wallstreetbets”

This figure plots the dynamic treatment effects between three months prior to the treatment and three months after the treatment to examine whether the “parallel trend” assumption holds for the “difference-in-differences” analysis on whether the first appearances of stock tickers on “Wallstreetbets” can have a positive and significant effect on stock crash risk. The “difference-in-differences” specification is as follows:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where the dummy variables $D_{i,j,c,t}$ indicate whether a stock i is treated in cohort c at time t, and the distance $j \in [-3, 3]$ from the current month to the treatment month. Month -1 is chosen to be the base month that will be omitted from the regression. Month 0 is the treatment month, and a green dotted line is plotted for better illustration. The coefficients on the rest of the dummies $D_{i,j,c,t}$ together with their 95% confidence interval bands are then plotted against their respective time periods. The blue markers display results using logit-generated crash risk as the dependent variable, while the golden markers use machine learning-generated crash risk. $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Standard errors are clustered at the unit level. The regression results are reported in Column (3) and (6) in Table 1.6.

because of their distorted beliefs (Brunnermeier et al., 2007). The combination of these factors should lead to a natural hypothesis that retail attention should have an outsized impact on the crash risk of smaller stocks and stocks with lower institutional ownership.

To examine this hypothesis, I divide the universe of stocks into two subgroups based on either lagged size or institutional ownership. Then I define a dummy variable $D_{size/io} = 1$ if the stock is larger than the median or zero otherwise, based on the lagged value of each stock three months prior to each event. In the case of institutional ownership, $D_{size/io} = 1$ if the ratio of institutional ownership for the stock is greater than the median or zero otherwise, based on the lagged value of institutional ownership three months prior to each event. Then I interact $D_{size/io}$ with the $Treated$ dummy variable in the same "stacked difference-in-differences" specification:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta_1 D_{i,c,t} + \beta_2 D_{i,c,t} \times D_{size/io} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \quad (1.7)$$

I report the results of this specification in Table 1.7. Columns (1) to (4) report the results of using logit-generated crash risk as the dependent variable, while Columns (5) to (8) use machine learning-generated crash risk as the dependent variable. Columns (1), (3), (5), and (7) only include the treated dummy and the interaction between the treated and the size dummy or IO dummy. Columns of even numbers add control variables. Standard errors are clustered at the unit level.

Table 1.7 First Appearances of Stock Tickers on “Wallstreebets” and Crash Risk: Size & IO

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Crash Risk (Logit)				Crash Risk (EEC)			
Treated	1.501***	1.038***	1.539***	0.988***	1.060***	0.434***	1.019***	0.422***
	(0.182)	(0.326)	(0.177)	(0.310)	(0.090)	(0.142)	(0.090)	(0.145)
Treated×D_{size}	-0.930***	-0.743**			-0.766***	-0.202		
	(0.205)	(0.343)			(0.104)	(0.153)		
Treated×D_{io}			-1.082***	-0.689**			-0.735***	-0.191
			(0.202)	(0.330)			(0.102)	(0.155)
Controls	NO	YES	NO	YES	NO	YES	NO	YES
Observations	208,502	125,734	208,502	125,734	208,502	125,734	208,502	125,734
R-squared	0.874	0.909	0.874	0.909	0.921	0.946	0.921	0.946
Cohort×Units FE	YES	YES	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES	YES	YES

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

This table reports results from a “stacked difference-in-differences” approach that examines whether the effect of first appearances of stocks tickers on “Wallstreetbets” on their ex-ante crash risk differs because of size or level of institutional ownership. “Wallstreetbets” was started in April 2012. From the beginning of “Wallstreetbets” to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of “never treated” stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the “cohorts” containing treated and control observations are stacked together and the following specification is run:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta_1 D_{i,c,t} + \beta_2 D_{i,c,t} \times D_{size} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $D_{size/io}$ is a dummy variable that equals one if the stock is larger than the median or zero otherwise, based on the lagged value of each stock three months prior to each event. In the case of institutional ownership, $D_{size/io} = 1$ if the ratio of institutional ownership for the stock is greater than the median or zero otherwise. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β_2 is the coefficient of interest that estimates the difference in average treatment effect on the treated stocks if the stocks belong to the large stock subgroup. Column (2) adds control variables that include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Standard errors are clustered at the unit level to account for possible duplicate observations.

Consistent with our hypothesis, the coefficient on the interaction term between the treated and size dummy or the IO dummy is negative and economically, and statistically significant. For example, as shown in Column (1) when controls are not included, if the stock is below median size, the first appearance on "Wallstreetbets" increases stock crash risk by 1.5 percentage points, much higher than our baseline estimate of 1.03%. If the stock is above the median size, the effect is much smaller at approximately 57 bps. In column (2) when control variables are included, being a small stock that first appears on 'Wallstreetbets" leads to a 1.04 percentage points increase in crash risk. The interaction term between $Treated$ and the size dummy remains significantly negative. The results are consistent when using institutional ownership as the main variable of interest. These results are consistent with prior literature that retail investors have a higher impact on smaller stocks or stocks with a lower level of institutional ownership.

1.6.4 Supporting Evidence from Trading Volume and Volatility

One necessary assumption for our analysis is that retail investors pile in the stocks that are mentioned on social media. While we do not have individual trading data, there should be a surge in trading volume and volatility (Foucault et al., 2011) around the events. To examine whether this is the case, we re-run the "difference-in-differences" analysis but substitute the dependent variable with trading volume and return volatility, where trading volume is defined as the monthly total volume of shares traded scaled by total shares outstanding, and volatility is defined as daily return volatility of the current month. The results are reported in Table 1.8.

The table shows clearly that there is a significant surge in both trading volume and return volatility in the treated stocks that first appeared on "Wallstreetbets". Moreover, the dynamic tests confirm that there is no evidence that the "parallel trend" assumption is violated. In fact, before the event happens, there is a downward trend for the treated stocks in terms of trading volume and return volatility. This can be more readily shown in Figure 1.5.

Taken together, these results support our main analysis that heightened retail attention as a result of social transmission leads to higher ex-ante crash risk. Moreover, there is evidence that retail activities are behind the surge of trading interests in these stocks.

Table 1.8 First Appearances of Stock Tickers on “Wallstreebets”: Trading Vol & Volatility

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES		Trading Volume			Volatility	
Treated	0.227***	0.146***		0.260***	0.162***	
	(0.026)	(0.032)		(0.021)	(0.026)	
Month -3			-0.067*			-0.037
			(0.034)			(0.041)
Month -2			-0.059*			-0.068*
			(0.031)			(0.038)
Month 0			0.336***			0.497***
			(0.042)			(0.042)
Month +1			0.066*			0.024
			(0.038)			(0.039)
Month +2			0.015			0.013
			(0.041)			(0.042)
Month +3			-0.017			-0.055
			(0.044)			(0.042)
Observations	209,478	125,748	125,748	212,961	125,748	125,748
R-squared	0.931	0.954	0.954	0.790	0.842	0.843
Cohort×Units FE	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES

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

This table reports results from a “stacked difference-in-differences” approach (Gormley and Matsa, 2011) that examines the effect of first appearances of stocks tickers on “Wallstreetbets” on their trading volume and volatility. Columns (1) to (3) use trading volume as the dependent variable, while Columns (4) to (6) use return volatility. “Wallstreetbets” was started in April 2012. From the beginning of “Wallstreetbets” to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of “never treated” stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the “cohorts” containing treated and control observations are stacked together and the following specification is run:

$$TradingVol_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $TradingVol_{i,c,t}$ is the trading volume of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Columns (1), (2), (4), and (5), where Columns (2) and (5) add control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity, MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Columns (3) and (6) examine the dynamic treatment effects around the events. Standard errors are clustered at the unit level.

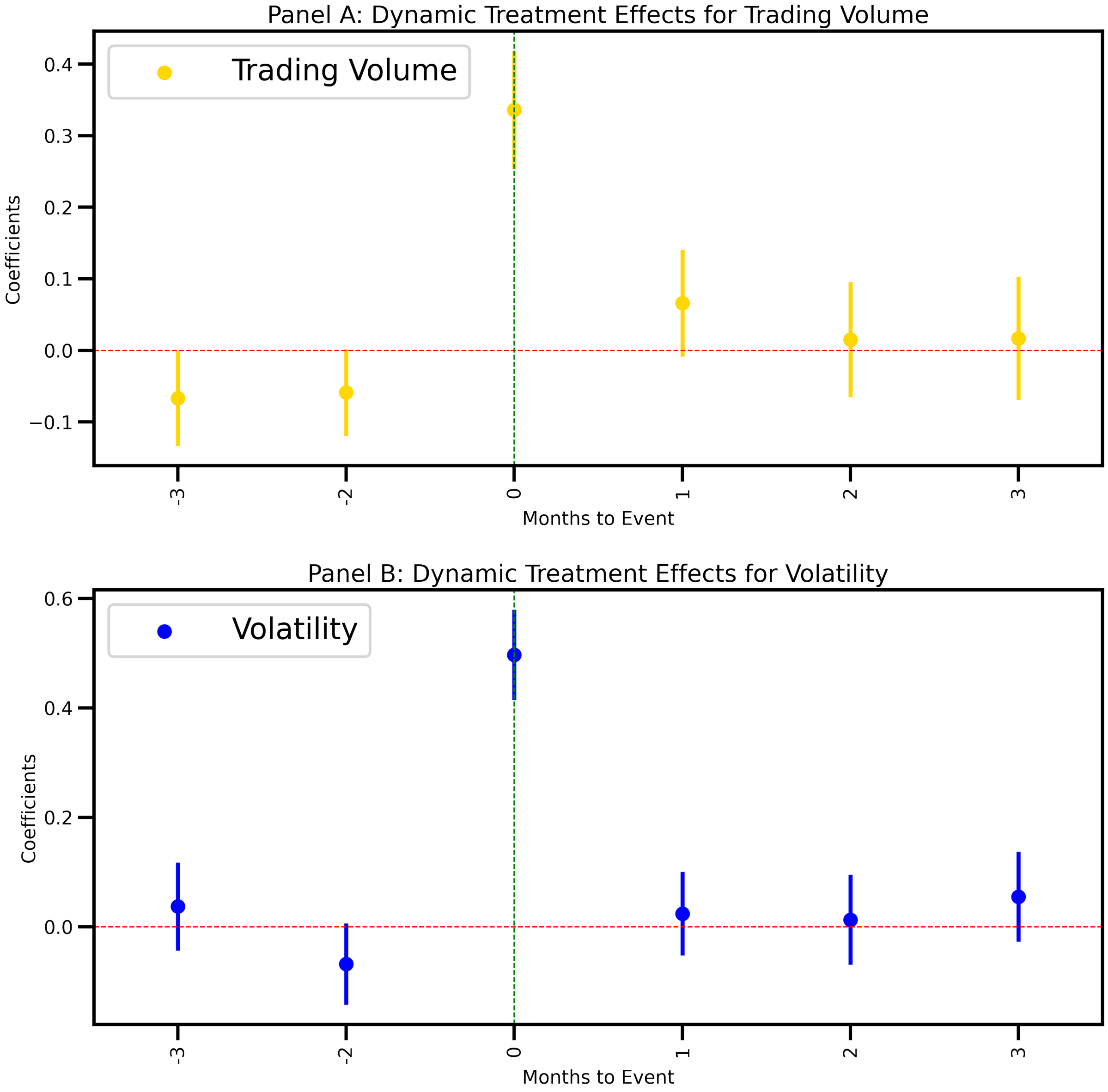


Figure 1.5 Dynamic Treatment Effects: Trading Vol & Volatility

This figure plots the dynamic treatment effects between three months prior to the treatment and three months after the treatment to examine whether the "parallel trend" assumption holds for the "difference-in-differences" analysis on whether there is a surge in trading volume and volatility after the first appearances of stock tickers on "Wallstreetbets". The "difference-in-differences" specification is the same specification as in our main test except that we replace the dependent variable with either trading volume or return volatility. Panel A display results on trading volume, while Panel B shows results for return volatility. The regression results are reported in Column (3) and (6) in Table 1.8.

1.7 Retail Traders and Crash Risk: Daily Evidence

In this section, we approach the main questions using the daily data by exploring the $SKEW$ measure by Xing et al. (2010), which is widely used as a proxy for firm-level crash risk (Bollen and Whaley, 2004; Van Buskirk, 2011; Kim and Zhang, 2014; Kim et al., 2016). It is motivated by the notion that a volatility smirk indicates investors' expectation of a steep decline in the underlying asset value (Bates, 2000).

Using $SKEW$ as a proxy has the following advantages. First, it is available at a daily frequency for stocks that have options traded. Second, it is easy to compute as it only relies on implied volatility. Third, it is ex-ante in nature and thus conforms to our purpose. Formally, $SKEW$ is defined as follows:

$$SKEW_{i,t} = ImpliedVol_{i,t}^{OTM-Put} - ImpliedVol_{i,t}^{ATM-Call} \tag{1.8}$$

Following Xing et al. (2010), I screen the options based on the following criteria. Days to expiration are between 10 and 60 days. Implied volatilities are between 0.03 and 2. Open interest must be greater than zero. Option price must be greater than \$0.125. Volume is non-missing. For out-of-money put options, the moneyness is between 0.8 and 0.95. For at-the-money call options, the moneyness is between 0.95 and 1.05. We choose the implied volatility of the put option with moneyness closest to 0.95, and the implied volatility of the call option with moneyness closest to 1 to compute the $SKEW$ measure for the day.

1.7.1 SKEW and Daily Returns

Xing et al. (2010) show that $SKEW$ is significantly negatively correlated with future weekly returns. To test whether this is the case in the daily frequency and to check whether daily $SKEW$ can be used as a suitable proxy for ex-ante crash risk, we need to examine whether $SKEW$ is significantly negatively correlated with future daily returns.

Therefore I follow Hu et al. (2021) and use the following specification:

$$R_{i,t} = \alpha + \beta SKEW_{i,t-1} + \sum_{p} \beta_p Control_{i,p,t-1} + \lambda_t + \epsilon_{i,t} \tag{1.9}$$

Where t is at a daily frequency. The control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. For robustness, I run both Fama-MacBeth regressions and panel regressions and report the results in Panel A of Table 1.9.

Panel A of Table 1.9 shows that throughout all specifications, the $SKEW$ measure is negatively correlated with future daily stock returns, which is statistically significant at the 1% level. These results corroborate the findings in the prior literature and provide support for using $SKEW$ as a valid proxy for ex-ante crash risk at the daily frequency.

1.7.2 Retail Trading of SKEW

In section 1.5, we show that retail investors have a tendency to buy high ex-ante crash risk stocks. To see whether this is also the case in the daily frequency, we again use the trading measure derived from Robintrack to regress on the contemporaneous $SKEW$ measure and the same set of control variables that we used in the previous test. We regress retail trading measures on the contemporaneous $SKEW$ measure instead of the lagged measure because we want to examine retail trading behavior on the "ex-ante" measure of crash risk. The results are reported in Panel B of Table 1.9.

To control for common market-wide shocks, we follow the prior specifications and include day fixed effects and cluster standard errors at the stock level. From Panel B of Table 1.9, we see that both regressions using different measures for retail trading load positively and significantly on the contemporaneous $SKEW$, the proxy for ex-ante crash risk measure. These results are consistent with our prior monthly results that retail investors tend to overbuy high crash-risk stocks.

Apparently, these results only report the positive correlation between crash risk and retail trading, while the causality can go both directions, just like in the monthly case. To see whether retail behaviors have a real influence on ex-ante crash risk, we turn to online conversations in "Wallstreetbets" again but follow a different path. We want to examine whether the intensity of

Table 1.9 Daily Returns, Retail Trading, and Crash Risk ($SKEW$)

	(1)	(2)	(3)	(4)
Panel A: Daily Stock Returns and Crash Risk ($SKEW$)				
VARIABLES	FMB		Panel	
Lag Option SKEW	-0.001***	-0.002***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Controls	NO	YES	NO	YES
Observations	2,071,209	2,010,815	2,071,209	2,010,815
R-squared	0.003	0.072	0.199	0.201
Panel B: Robinhood User Trading and Crash Risk ($SKEW$)				
VARIABLES	Change in Log(Robinhood Users)		% Change in Robinhood Users	
Option SKEW	0.001**		0.001**	
	(0.000)		(0.001)	
Controls	YES		YES	
Observations	703,614		862,423	
R-squared	0.011		0.003	
Note:			*p<0.1; **p<0.05; ***p<0.01	

This table examines the relationship between daily returns and lagged $SKEW$ measure, and the relationship between Robinhood user trading and the contemporaneous $SKEW$ measure. Panel A reports regressions of daily stock returns on lagged $SKEW$ measure as a proxy for crash risk in the daily frequency. The $SKEW$ measure follows Xing et al. (2010):

$$SKEW_{i,t} = ImpliedVol_{i,t}^{OTM-Put} - ImpliedVol_{i,t}^{ATM-Call}$$

The option data is from Option Metrics. We screen the option data based on the following conditions. Days to expiration are between 10 and 60 days. Implied volatilities are between 0.03 and 2. Open interest must be greater than zero. Option price must be greater than $0.125. Volume is nonmissing. For out-of-money put options, the moneyness is between 0.8 and 0.95. For at-the-money call options, the moneyness is between 0.95 and 1.05. We choose the implied volatility of the put option with moneyness closest to 0.95, and the implied volatility of the call option with moneyness closest to 1 to compute the $SKEW$ measure for the day. Columns (1) and (2) report Fama-MacBeth cross-sectional regressions, while Columns (3) and (4) report panel regressions. The control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. For panel regressions, we include day fixed effects, and standard errors are clustered at the stock level. Panel B reports panel regressions of Robinhood user trading measures on contemporaneous $SKEW$ measure as a proxy for ex-ante crash risk and control variables. The trading measures include the change in the log of user numbers and the percentage change of user numbers from the previous day.

daily conversations about certain stocks can have a significantly positive impact on the ex-ante crash risk of these stocks.

1.7.3 Online Conversations and SKEW: Endogeneity

Apparently, online conversations about stocks are endogenous. As shown in Han et al. (2022), agents receive prominent presentations of other agents' trading strategies, typically represented by high past returns, and thus follow the same strategy, which leads to feedback on the stock returns. Because of this feedback loop, it's impossible to separate the two legs of the circle via the usual regression specifications.

Specifically, consider the following specification, where we regress the $SKEW$ measure on the number of times each stock is mentioned on social media, controlling for a set of stock characteristics.

$$SKEW_{i,t} = \alpha_0 + \beta SocialTransmission_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + \sigma_i + \epsilon_{i,t} \tag{1.10}$$

In a slight abuse of notation, the $t-1$ in the subscript of "Social Transmission" means the pre-trading hours from 16:30 PM on the previous day to 09:00 AM on the current day, while the $t-1$ in the controls ranges from the previous day to previous month, depending on the variable referred to. In this specification, even when we use the two-way fixed effects estimator, "Social Transmission" is still correlated with the idiosyncratic error term, and thus the estimate of the coefficient β is inconsistent.

1.7.4 A Plausible Instrument

Let's consider the following scenario. Person A zones away during his long and boring working hours by wandering aimlessly on social media. His/her favorite venue for wandering is Reddit, a popular platform for talking about anything. Each sub-venue specializing in a different topic is called a "Subreddit", a symbol of rich social life in a society. Apart from working, person A spends a tremendous amount of time on hobbies such as football, fishing, and political debates, where he/she posts and comments on the corresponding Subreddits. Apart from all this, person A has developed a keen interest in stock trading, and thus becomes a subscriber of "Wallstreetbets", as he/she can always find interesting ideas for trading there. For person A, Reddit almost satisfies all his/her needs for socializing, and the migration cost is high, plus there is no comparable platform (Chang et al., 2014).

Therefore, person A's activities on "Wallstreetbets" are correlated with his/her activities on other Subreddits. In other words, person A is more likely to post on "Wallstreetbets" if he/she is also posting on other Subreddits. However, it is logical that person A's activities on other Subreddits have no direct bearing on stock market returns. Such an influence can only be exerted via his/her activities on "Wallstreetbets".

Formally, consider the following specification.

$$WSB\_Posts_{i,t-1} = \alpha_0 + \beta_Z Non\_Finance\_Posts_{i,t-1} + \epsilon_{i,t-1} \tag{1.11}$$

$$SKEW_{i,t} = \alpha_1 + \beta_X WSB\_Posts_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + u_{i,t} \tag{1.12}$$

Where the first equation represents the first-stage regression, and the second equation represents the instrumental variable estimation. The subscript i represents stock i and simultaneously all the agents that mention stock i. To operationalize this procedure, we must ensure that the non-finance conversations are truly non-finance related. Therefore, the name of the Subreddit matters.

To ensure that we extract non-finance posts from non-finance "Subreddits", I follow the strategy used in Li et al. (2021). I choose a set of "seed words" and find out 50 words/phrases that are closest in meaning to each seed word. Finally, I choose those "Subreddits" whose title does not contain these keywords. The seed words I choose include: "finance", "stock-market", "stocks", "wall-street", "trading", "forex", "options", "investment", "bond-market", and "bonds".

How to find words/phrases that are similar in meaning to the seed words? Recent advances in computational linguistics offer powerful tools to help solve the problem. First, we want to vectorize the words/phrases into fixed-length vectors. Then we compute the cosine similarity between each pair of vectors to check their distance to each other as a proxy for meaning closeness. To do this, I use the pre-trained word embedding system called "Global Vectors for Word Representation" (GloVe) developed by Pennington et al. (2014). These vectors are trained on the whole corpus of

Wikipedia up to 2014 and Gigaword 5 (Parker et al., 2011) on the co-occurrences of words and phrases. I use the 300D version of GloVe, which means that each word/phrase is represented by a 300 dimension vector: $V = [x_1, x_2, ..., x_{300}]$. Thus the cosine similarity between two words V_1 and V_2 is:

$$CosineSim_{1,2} = \frac{V_1 \cdot V_2}{||V_1|| \cdot ||V_2||} \tag{1.13}$$

The cosine similarity measure for word vectors ranges between zero and one, with one being the closest meaning.[6] I find out the top 50 most similar words/phrases for each seed word and group them together. Because of duplicates, we end up with a set of 351 keywords that are related to the topic of finance. I then use these keywords to screen all the Subreddits.

1.7.5 Instrumental Variable Results

With all the data processed, we are ready to construct the instruments. First, we denote a user j's number of posts on "Wallstreetbets" about stock i on day t as $n_{i,j,t}^{WSB}$, and his/her number of posts on non-finance "Subreddits" as $n_{i,j,t}^{nonFin}$. Then stock i's total number of posts on "Wallstreetbets" on day t is $N_{i,t}^{WSB} = \sum_j n_{i,j,t}^{WSB}$. The instrument we construct for this variable would be $N_{i,t}^{nonFin} = \sum_j n_{i,j,t}^{nonFin}$, where the term is summing over all j that have posted on "Wallstreetbets" about stock i on day t.

We proceed to run the regressions of the daily $SKEW$ measure on our main variable of interest – the number of "Wallstreetbets" posts $N_{i,t}^{WSB}$, instrumented by the total number of non-finance posts by the same users $N_{i,t}^{nonFin}$, controlling for the same set of independent variables we use in prior settings. First, we run panel regressions without using the instrument. Then, to test whether there is evidence that the instrument violates the exclusion restriction, I add the instrument into the regression to see whether the instrument is inappropriately excluded. Finally, I run the regression with instrumental variable estimation. The first stage regression is untabulated, but the coefficient on the instrument is 0.049 and statistically significant at the 1% level, and the R^2 is 3.4%. I report the main results in Table 1.10.

The insignificant coefficient on the number of non-finance posts in Column (2) supports the

[6] Because all word vectors contain nonnegative numbers, the cosine similarity between any pair of word vectors is nonnegative.

Table 1.10 Instrumental Variable Estimation: “WSB” Posts and Crash Risk ($SKEW$)

VARIABLES	(1) Panel	(2) Panel	(3) IV
Number of “Wallstreetbets” Posts	0.070***	0.067***	0.193***
	(0.019)	(0.018)	(0.035)
Number of Non-Finance Posts		0.005	
		(0.004)	
Controls	YES	YES	YES
Observations	2,655,209	2,655,209	2,655,209
R-squared	0.089	0.089	0.042
Day FE	YES	YES	YES
Firm Cluster	YES	YES	YES
Note:			*p<0.1; **p<0.05; ***p<0.01

This table reports the results of regressing the daily $SKEW$ measure on the number of “Wallstreetbets” posts, controlling for other stock characteristics. Column (1) reports a panel regression of $SKEW$ on the number of “Wallstreetbets” posts. Column (2) adds the proposed instrument “number of non-finance posts” to test the exclusion restriction. Column (3) reports the result of instrumental variable estimation. Denote a user j's number of posts on “Wallstreetbets” about stock i on day t as $n_{i,j,t}^{WSB}$, and his/her number of posts on non-finance “Subreddits” as $n_{i,j,t}^{nonFin}$. Then stock i's total number of posts on “Wallstreetbets” on day t is $N_{i,t}^{WSB} = \sum_j n_{i,j,t}^{WSB}$. The instrument we construct for this variable would be $N_{i,t}^{nonFin} = \sum_j n_{i,j,t}^{nonFin}$, where the term is summing over all j that have posted on “Wallstreetbets” about stock i on day t. The IV specification is as follows:

$$N_{i,t-1}^{WSB} = \alpha_0 + \beta_Z N_{i,t-1}^{nonFin} + \epsilon_{i,t-1}$$

$$SKEW_{i,t} = \alpha_1 + \beta_X N_{i,t-1}^{WSB} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + u_{i,t}$$

The $t-1$ subscripts on the number of posts refer to the time period of 16:30 PM the previous day to 9:00 AM on day t. The first stage regression of $N_{i,t}^{WSB}$ on $N_{i,t}^{nonFin}$ produces a coefficient of 0.049, statistically significant at the 1% level, and a R^2 of 3.4%, which dispels the weak instrument concern. In all specifications, the control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. We include day fixed effects, and standard errors are clustered at the stock level.

exclusion restriction assumption. The significantly positive coefficient on the number of "Wallstreetbets" posts in the instrumental variable estimation in Column (3) is consistent with our prior results of the "Difference-in-Differences" specification that online conversations among retail investors positively influence the ex-ante crash risk of stocks. A one-standard-deviation increase in the number of "Wallstreetbets" posts is associated with a 15 bps increase in the $SKEW$ measure on average. Since the mean $SKEW$ is 0.065, the 15 bps increase translates into approximately 2.3% increase in ex-ante crash risk on a daily basis.

These results, combined with our prior results on monthly crash risk, support our hypothesis that social media conversations are instrumental in facilitating more efficient herding of individual investors, which in turn drives the increase in the ex-ante crash risk of the underlying stocks.

1.8 Conclusion

Recent development in financial technology (FinTech) like "Robinhood" has dramatically reduced the hurdle for retail trading. In addition, popular online forums like "Reddit" facilitate more efficient sharing of trading ideas. These innovations can likely amplify the effect of correlated retail trading behaviors. Because of distorted beliefs, retail investors tend to over-buy high crash-risk stocks, contributing to the negative price of crash risk. The buying activities and subsequent price reactions formulate a possible feedback loop. The resulting more elevated level of crash risk contributes to exacerbated market volatility, potentially damaging investor welfare.

Future research avenues could further explore social media's role in forming investor beliefs and their subsequent trading behavior. As reflected in the meme stock frenzy, the mass psychology of the online investing community could be influenced without apparent fundamental information, often to the harm of such investors. Studying this interaction between social media conversations and asset prices could help us understand the intricacies of price formation and aid policymakers in their pursuit of protecting potentially novice and vulnerable investor groups.

CHAPTER 2

THE CYBER RISK PREMIUM

2.1 Introduction

The digital transformation of the economy and increased interconnectivity have created unprecedented opportunities and under-explored risks. On the one hand, the ever-growing pool of data along with advances in artificial intelligence helps to promote efficiency and productivity in the economy; on the other hand, it exposes households, businesses, and governments to a new and potentially systemic source of risks: the cyber risk.[1] In the past decade, various forms of cyberattacks have attracted widespread attention. For instance, the Equifax data breach in 2017 exposed approximately 147 million names and dates of birth, 145.5 million Social Security numbers, and 209,000 payment card numbers along with expiration dates, which led to a settlement of nearly $700 million with federal and state investigators. The recent SolarWinds cyberattack, which came to light in December 2020, illustrates both the vulnerability of the most secure networks and the potential for immense collateral damage to corporate and governmental entities connected to these networks.

Considering the importance of cyber risk, it is natural to study how it might influence stock market prices and returns. The primary challenge for such a study is that cyber risk is latent, not directly observable. To address this challenge, we use machine learning algorithms to develop a real-time estimate of the likelihood for each individual firm to experience a cyberattack in the subsequent period. The underlying logic is that hackers do not choose their targets at random, but focus on firms with certain attributes (see, e.g., Kamiya et al. (2021)). Moreover, firms tend to communicate their self-perceived exposures to cyber risk through their disclosures, e.g., through 10-K filings. This information is also likely to be important in predicting the level of the firm's cyber risk. Machine learning techniques are particularly suited to this task due to their ability to extract useful information from a large set of features and to deliver superior out-of-sample forecasts.

[1] See Kashyap and Wetherilt (2019) for discussions on the distinguishing features of cyber risk.

In our empirical analyses, we introduce a novel data set of cyberattacks and apply a variety of machine learning techniques including the logistic ridge regression, the K-Nearest Neighbor, and Naive Bayes combined with an “EasyEnsemble" sampling technique (EEC-KNN and EEC-NB).[2] We find that our cyber risk measure based on these algorithms has a superior ability to forecast the occurrence of future cyberattacks. For instance, compared with the simple logistic forecasting model, the predictive performance of the logistic ridge regression improves from 0.4% to 6.3% based on the harmonic mean of precision and recall rates (F-score) and from around 1.4% to 59.9% based on the geometric mean of sensitivity and specificity (G-mean). Because the different machine learning techniques yield similar results, we present our main results using the logistic ridge regression, which is commonly used in asset pricing literature.

Armed with the cyber risk measure, we study how cyber risk is related to stock returns. Our primary tests use the Fama and MacBeth (1973b) cross-sectional regressions to examine the incremental predictive power of the cyber risk measure for stock returns against well-known firm characteristics in the period from July 2008 to June 2019. In the first set of regressions, we include the natural log of market capitalization, book-to-market ratio, gross profitability, investment ratio, and past 11-month return skipping the most recent month as control variables. In the second set, we add past one-month return, idiosyncratic volatility (Ang et al., 2006), and the Amihud illiquidity ratio (Amihud, 2002). Finally, we also control for the organizational capital (Eisfeldt and Papanikolaou, 2013), CAPM market beta, tail risk beta (Kelly and Jiang, 2014), co-skewness (Harvey and Siddique, 2000), and net operating assets (Hirshleifer et al., 2004). In a contemporary study, Florackis et al. (2022) propose a cyber cosine measure based on the linguistic similarity between hacked and non-hacked firms in the risk factor section of their annual reports, which is also included in some of our regression specifications. In all these regressions, the relation between cyber risk and stock returns is economically meaningful and statistically significant. In terms of magnitudes, a one standard deviation increase in cyber risk is associated with higher average

[2] The EasyEnsemble technique combines undersampling and bootstrapping, which is effective in dealing with class-imbalance problems. This technique is particularly suited to address issues like cyberattacks, which represent rare events in the data. See Section **??** for relevant technical details on the EasyEnsemble technique.

stock returns of between 0.16% and 0.21% per month. These results show that cyber risk is an independent and important driver of cross-sectional variation in stock returns during our sample period.

We also perform a standard portfolio analysis. At the end of each June from 2008 to 2018, we sort stocks into five quintiles based on the estimated level of cyber risk using information available in real-time and form value-weighted portfolios that are rebalanced at the end of the following June. Consistent with the Fama-MacBeth regressions, we find that stocks in the top quintile with the highest level of cyber risk have higher returns than those in the bottom quintile. The spread in returns is 0.84% per month after we adjust for their market exposure and 0.58% per month after adjusting for Fama-French five factors and momentum. The alpha remains large and statistically significant when we form tercile and quartile portfolios as well as under alternative asset pricing models.

In addition to the cross-sectional variation in average returns, we examine the time variation in the relative returns between stocks with high and low cyber risk. We conduct two tests. First, we study the relation between the variation in the return difference between high and low cyber risk stocks and the New York University's Index of Cyber Security (ICS), which is based on monthly surveys of industry executives about their perceived level of cyber risk. Because ICS is a monthly aggregate measure of cyber security, we form a factor-mimicking portfolio and compute its performance. We estimate each stock's return sensitivity to changes in ICS in the past year. A lower ICS beta is associated with higher firm-level cyber risk. Then we form a spread portfolio that buys stocks with high ICS beta and sells those with low ICS beta. The return on the spread portfolio based on our cyber risk measure and that based on the ICS beta have a strong negative comovement, with a time-series correlation coefficient of −32%.

Second, we study the relation between the performance of the spread portfolio based on our cyber risk measure and that of two cybersecurity ETFs that invest in firms providing cybersecurity services. We again find strong negative comovements between the two, with time-series correlation coefficients below −40%. That is, when stocks with higher cyber risk underperform, cybersecurity firms tend to have higher returns. These results suggest that the time variation in the return spread between high and low cyber-risk firms is likely driven by the market's perception of cyber risk in the economy.

The strong relation between our cyber risk measure and future stock returns is consistent with the view that cyber risk is priced in the stock market. We conduct a number of tests to strengthen the identification of the cyber risk premium. First, we exploit the variation along the dimension of industry competition. Kamiya et al. (2021) shows that the negative stock market reaction to cyberattacks on victim firms spills over to their peers around the announcement of cyberattacks. We find that this spill-over effect is larger when the victim firm is a stronger competitor of its peer firms in the product market. This is because weaker peer firms are less likely to be able to increase market share at the expense of a stronger competitor—the repricing effect of their shares due to market recognition of their cyber risk exposure is less confounded by the effects of product market competition. Second, firms that provide products and services similar to those of hacked firms are likely to hold similarly valuable data, which makes them vulnerable to cyberattacks. We use the product similarity measure proposed by Hoberg and Philips (2016) as a proxy for data similarity. We find that peer firms with higher data similarity to victim firms tend to experience a more negative stock market reaction around the announcement of cyberattacks, and that data similarity is a positive contributor to our cyber risk measure. These tests provide further evidence for cyber risk being an important determinant of stock returns.

We evaluate the robustness of our results by performing additional tests. First, we split our sample into two subperiods. We find that in both periods, the cyber risk measure is positively related to future stock returns, with the effect slightly stronger in the more recent subperiod. Second, we construct a cyber risk measure based on industry-adjusted risk disclosure variables and compute industry-adjusted returns for each stock in our sample. In both cases, we find that the positive relation between cyber risk and future stock returns remains strong. Third, to address the concern that our cyber risk measure may be driven by other dimensions of risks such as corporate fraud, we use our measure to predict the incidence of corporate misconduct and financial misconduct. Our tests empirically reject this hypothesis. Fourth, we consider different machine learning algorithms and use different dictionaries in linguistic analyses to estimate the cyber risk measure. The results are robust. Finally, we consider a placebo test to pinpoint the importance of cyber risk. In particular, we randomly select firms to construct a fake sample of cyber-attacked firms. The number of random draws to create the fake sample equals the actual number of cyberattacks for each industry in each year. We then use the same machine learning algorithms to estimate each firm's pseudo-cyber risk and estimate its relation to stock returns. Our results show that this relation is essentially flat. This result highlights the importance of cyber risk in driving stock returns.

Our study contributes to the fast-growing literature that studies the implications of cyber risk for firms, the financial markets, and the economy. Many studies in this literature focus on the impact of realized cyberattacks. For instance, Kamiya et al. (2021) find that the announcement of successful cyberattacks is, on average, associated with a 1.09% wealth loss for shareholders within a three-day window around the incidents. They argue that successful cyberattacks lead to value loss for victims for both actual and reputational reasons.[3] Notable exceptions include Jamilov et al. (2020), and Florackis et al. (2022), who use textual information to identify firms with high cybersecurity risk. Jamilov et al. (2020) use firms' earnings conference call transcripts to develop a measure for firm-level cyber risk exposure and sentiment and report a number of interesting findings, such as an increase in corporate discussions of cyber risk in earnings calls, increasingly negative sentiment regarding cyber risk, and the spread of cyber risk discussions across regions.[4] However, they do not distinguish between cyber risk exposure and cyber risk awareness. For instance, corporate managers' extensive cyber risk discussions in conference calls can be driven by their keen awareness of cyber risk; a lack of such discussions can be driven by

[3] In another study, Michel et al. (2020) study the timing of cyberattack announcements. They find negative abnormal returns on cyberattack victims prior to the announcement of the attacks, which is consistent with some information leakage. Echoing this message, Lin et al. (2020) reports evidence of insider trading ahead of public announcements of cyberattacks. Binfarè (2019) studies the effect of data breaches on the cost of debt financing. He finds that lenders tend to charge borrowing firms larger spreads after they experience data breaches.

[4] Kopp et al. (2017) and Warren et al. (2018) examine the impact of cyberattacks on financial institutions and financial market infrastructure, and argue such attacks potentially can be quite damaging. Duffie and Younger (2019) study the linkage between the cyber security of large banks and financial stability. They argue that large banks tend to have sufficient liquidity to weather relatively extreme cyber runs; however, severe cyberattacks may deter nonbanks from sending funds through these institutions, which could create systemic risk.

their inadequate attention to or even ignorance of cybersecurity risk. Our results are robust to excluding textual variables from the cyber risk measure. Closer to us, Florackis et al. (2022) study the relation between cyber risk and stock returns. They use firms' descriptions of cyber-related risk in 10-K reports to identify firms with high cyber risk. Their conclusion that cybersecurity risk is priced in the cross-section of stock returns provides independent support for our results. The main difference between their study and ours is methodological. Florackis et al. (2022) compare the word distribution of "Item 1A. Risk Factors" in the 10-K reports of training firms with that of the hacked sample, identifying the most similar firms as high cybersecurity risk firms. Their method is simple and powerful, relying only on the textual data in firms' annual reports. Our machine learning algorithms provide a mapping from any feature set to the occurrence of cyberattacks. The special strength of our approach is its flexibility: the feature set can be expanded to encompass any form of data, including traditional and alternative data.

Our paper also joins the burgeoning literature that applies machine learning techniques to asset pricing. This literature has focused on using machine learning algorithms to efficiently select and combine firm characteristics to predict stock returns (see, e.g., Freyberger et al. (2020); Gu et al. (2020)) and to construct a robust stochastic discount factor (see, e.g., Kelly et al. (2019), Kozak et al. (2020), and Chen et al. (2020)). Instead of directly targeting asset returns, our paper uses machine learning techniques to estimate an important yet latent risk as economies become more digital. Our approach exploits the correlations between firm characteristics and the likelihood of cyberattacks and shows that the resulting estimate of cyber risk has predictive power for stock returns beyond the usual firm characteristics.

Finally, our paper is related to the emerging literature that applies computational linguistics to finance. This literature has made substantial progress in building subject matter lexicons. For instance, Tetlock (2007) used lexicons that are outside the finance field to detect tones in newspapers and their implications for returns and trading volumes. Loughran and McDonald (2011) developed a lexicon that classifies finance-related words into different sentiment classes, and shows that applying this lexicon to 10K filings can predict subsequent returns. More recently, researchers have explored deeper structures of texts that link to capital market activities. For example, Cohen et al. (2020) use changes in 10K texts to predict return, earnings, and bankruptcies. Ke et al. (2019) use supervised learning to extract sentiments from newspapers to predict returns. Bybee et al. (2020) apply topic modeling (an unsupervised word cluster approach) to gauge the relationship between news article topics and macroeconomic activities. Our paper contributes to the literature by introducing dictionaries developed in the cyber risk community to estimate individual firms' self-assessment of cyber risk in their 10-K filings and using it as an input to our machine learning algorithms to better predict cyberattacks.

The rest of the paper is organized as follows. Section 2.2 describes the data. Section 2.3 shows the methodology to construct the cyber risk measure. Section 2.4 presents the key results on the relationship between cyber risk and stock returns. Section 2.5 provides further identification of the cyber risk premium. Section 2.6 concludes.

2.2 Data

2.2.1 Cyberattacks

In this paper, we use a novel data set obtained from the Identity Theft Resource Center (ITRC).[5] It has a number of advantages over the data set from the Privacy Rights Clearinghouse which has been used previously in the literature.[6] It provides more up-to-date data, reports many more cyberattacks on public firms, and importantly, contains the source of the reports, which allows for cross-validation. ITRC provides annual data breach reports from the year 2005, detailing all the reported and confirmed cyberattack incidents for US-based organizations. These reports are stored on their website in the form of PDF files. We obtained the available annual reports from 2005 to 2019 and extracted all items connected to cyber incidents. We first matched them to Compustat firms using a fuzzy matching algorithm, and then manually checked each one of the matched pairs

[5] *ITRC*, https://www.idtheftcenter.org/. On their website: "The ITRC is a non-profit organization established to support victims of identity theft in resolving their cases, and to broaden public education and awareness in the understanding of identity theft, data breaches, cyber security, scams/fraud, and privacy issues." See this SEC report that cites the data from ITRC: https://www.sec.gov/files/speech-jackson-cybersecurity-2018-03-15-data-appendix-updated.pdf.

[6] *Privacy Rights Clearinghouse*, 2019, https://www.privacyrights.org/data-breaches.

for confirmation. After this process, we are left with 1,010 cyberattack incidents.[7] Since we use the cyberattack as a binary variable, we count multiple attacks of a firm in a given year as one instance. This leaves us with 552 unique firm-year pairs, with 368 unique firms that experienced at least one cyberattack incident. Because we use accounting variables as predictors, we follow the convention in Fama and French (1996) and define each year as the 12 months from July to the following June. The sample then spans from 2006 to 2018.[8] Figure 2.1 plots the number of unique incidents and the unique firm-year pairs for firms that experienced cyberattacks during the sample period. It shows an increasing time trend in cyberattacks.

An important question is whether the distribution across industries is uniform, as some industries are more digitized than others, and for some, significantly more data exist. We plot the number of incidents for each of the Fama-French 17 industries per Fama and French (1988) in our sample in Figure 2.2.

From Figure 2.2, we find that the financial sector is the hardest-hit industry, with a total of 247 incidents, dwarfing all other sectors. This is not surprising because financial firms possess large amounts of client identification and financial data. With the exception of the "Mining and Minerals" industry, no sector appears immune to cyberattacks as firms increasingly rely on online platforms to conduct their business and on cloud services to store their data.

Before leaving this subsection, we shall note that in an important paper, Cong et al. (2023) point out that a large fraction of cybercrimes and cyberattacks are unreported and kept hidden by victim firms. The under-reporting and under-recording would lead researchers to miss instances of cyberattacks, the inclusion of which may increase the statistical power of empirical tests to identify the risk premium. One useful observation from our analyses is that when we restrict our sample of cyberattacks to significant incidents with strong stock market reactions, the algorithm shows stronger power in identifying the risk premium associated with cyber risk. If the incentive to underreport cyberattacks is stronger for more significant incidents, it would lead us to underestimate

[7] This increases our data set from the 311 observations from PRC, which is nearly a 3-fold increase.

[8] We drop the first 6 months of the calendar year 2006 and the last 6 months of the year 2019, as they do not constitute a whole year per our definition.

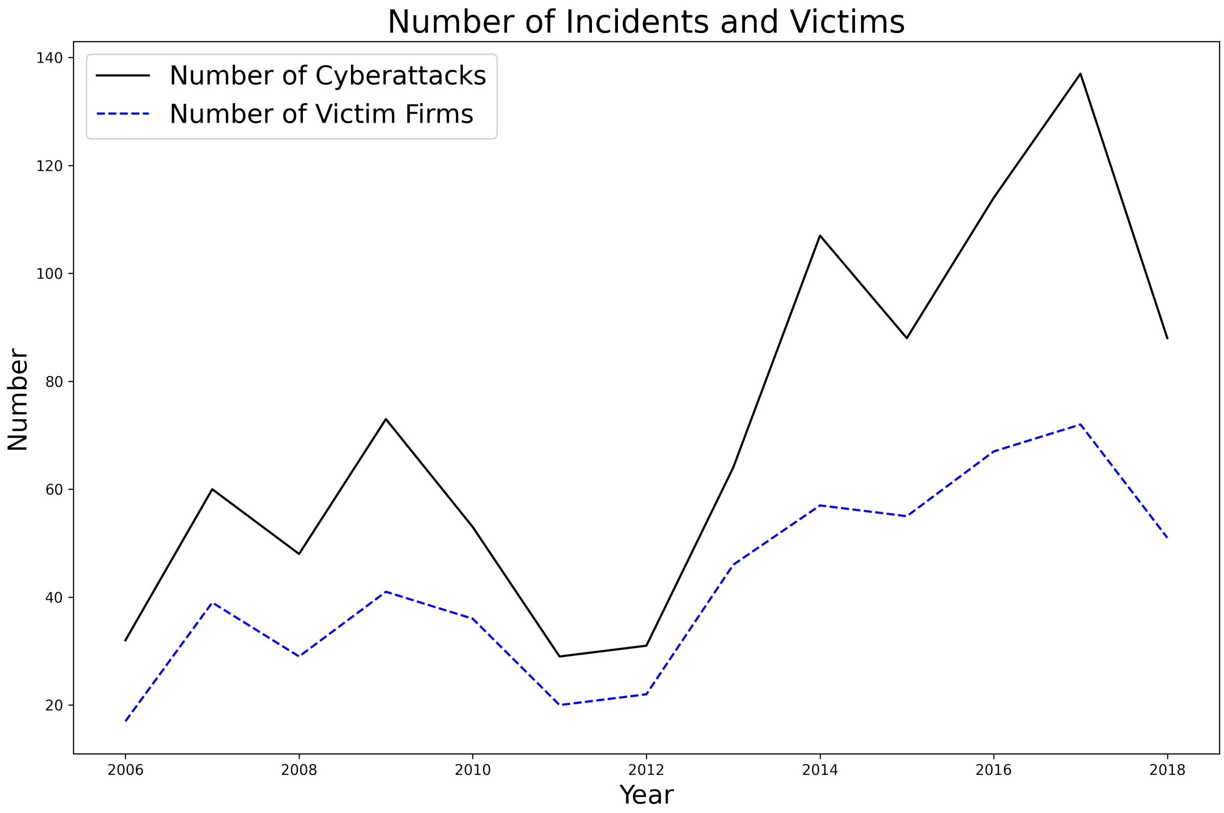


Figure 2.1 Disclosed Cyberattack Incidents for US Public Firms

The figure depicts the number of cyberattack incidents that are disclosed for US public firms from 2006 to 2018. The year convention follows Fama and French (1996), which starts from July to next June. The source of data is the Identity Theft Resource Center (ITRC). The black line indicates the number of unique cyberattack incidents matched to Compustat firms; the blue dashed line shows the number of unique public firms that experienced cyberattack incidents each year.

the importance of cyber risk in the stock market. Future research will benefit from a more comprehensive data coverage of cyberattacks.

2.2.2 Other Data

The rest of the data sources are as follows. Stock price and return data are from the CRSP. Accounting data are from Compustat. Asset pricing factors are from Kenneth French's website.[9] The 10-K filings are from the University of Notre Dame Repository website.[10]

[9] *French, Kenneth*, 2019, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[10] *McDonald, Bill, University of Notre Dame*, https://sraf.nd.edu/data/stage-one-10-x-parse-data/.

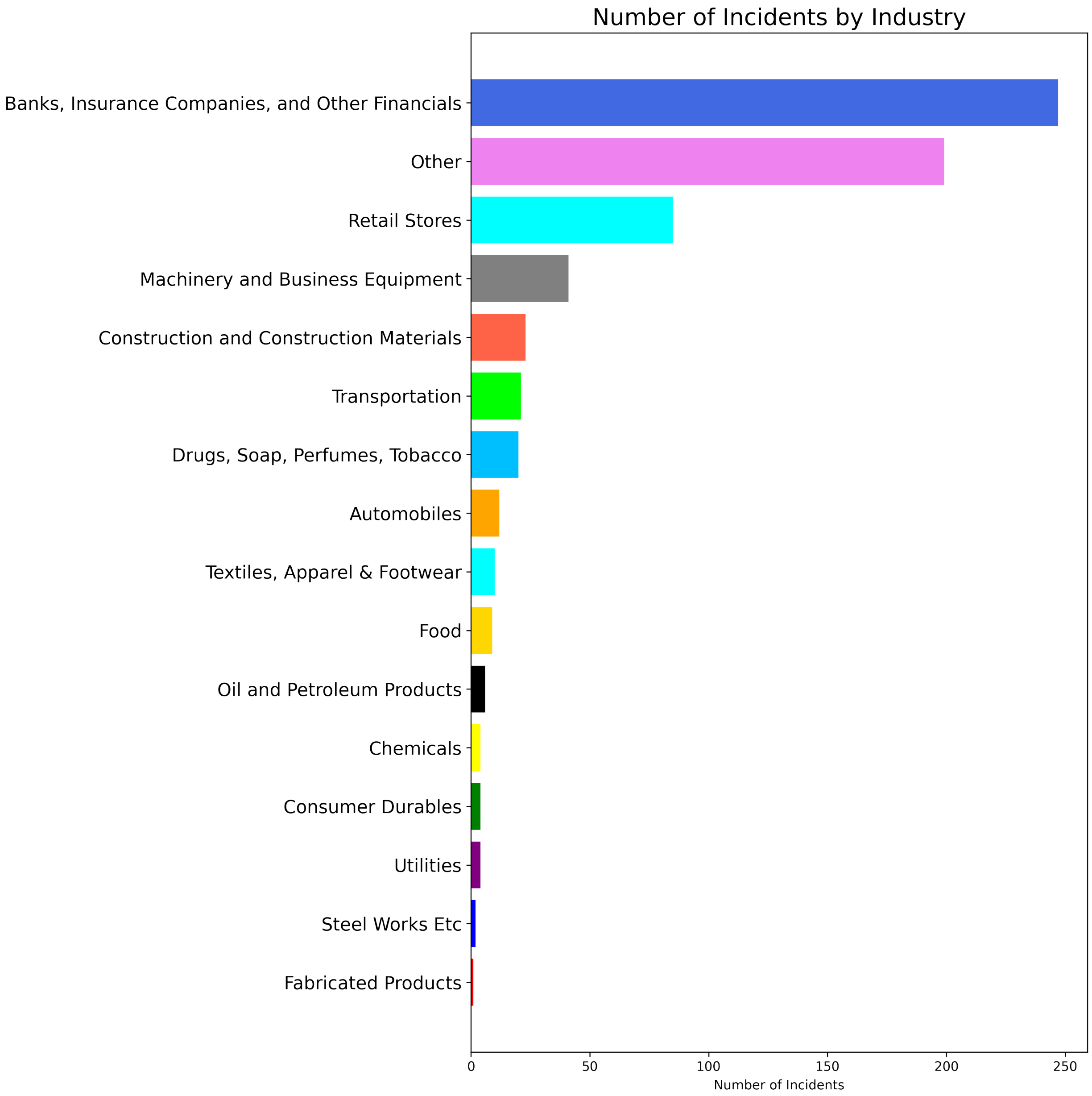


Figure 2.2 Industry Distribution of Cyberattacks

The figure depicts the industry distribution of attacked firms based on Fama-French 17 industry classification per Fama and French (1988). The bars show the number of incidents in the sample period in each industry. The hardest hit is the financial industry, with a total of 247 incidents. The least hit is the "Mining and Minerals" industry, which has zero incidents in our sample. The sample runs from 2006 to 2018.

2.2.3 Summary Statistics

Table 2.1 summarizes the characteristics of firms that experience cyberattacks (victims) and those that do not (non-victims) in our sample. All the variables are measured prior to the year when the incidents happen. The accounting variables are lagged by six months to ensure data availability. We follow Kamiya et al. (2021) in the definitions of the firm characteristics. In addition, we use two linguistic variables: the "Risk Factor Length," which is based on the length of "Item 1A: Risk Factors" in the firm's 10-K disclosure, and the "NIST Counts," which is the number of times a firm mentions a cybersecurity-related term in Item 1A, based on the NIST dictionary. We discuss this dictionary in more detail in Section 2.3.

Table 2.1 shows that victim firms tend to be larger, which suggests that hackers target firms with larger customer bases and bigger data sets, as such targets are likely to offer them potentially higher gains. They also tend to have more intangible assets, have higher cash flow, spend less on research and development, have a higher return on assets, and experience higher past year excess returns. These characteristic differences are in line with the results in Kamiya et al. (2021). In addition, victims tend to spend more time discussing cybersecurity-related risks in their 10-K disclosure. In the next section, we describe how we use these characteristics to construct the firm-level measure of cyber risk.

2.3 Measuring Firm-Level Cyber Risk

A well-known insurer, Northbridge Insurance, provides a broad definition of cyber risk: "Cyber risk commonly refers to any risk of financial loss, disruption or damage to the reputation of an organization resulting from the failure of its information technology systems."[11] To operationalize this notion, we follow the literature (e.g., Kamiya et al. (2021) and Michel et al. (2020)) and treat reported cyberattacks as realized cyber risk. Based on these cyberattacks, we use machine learning algorithms to develop a real-time estimate of the likelihood for each individual firm to experience a cyberattack in the subsequent period.

[11] *Northbridge Insurance*, 2019, https://www.nbins.com/blog/cyber-risk/what-is-cyber-risk-2/.

Table 2.1 Summary Statistics

Variable	Victim	Mean	STD	P25	P75
Intangibility	0	0.79	0.23	0.70	0.97
	1	0.82	0.21	0.74	0.97
CAPX/AT	0	0.04	0.06	0.01	0.05
	1	0.03	0.04	0.01	0.05
CF/AT	0	0.02	0.22	0.01	0.11
	1	0.07	0.09	0.02	0.12
Firm Size	0	6.68	2.03	5.26	8.02
	1	9.22	2.30	7.56	10.99
Net Worth	0	0.45	0.25	0.25	0.66
	1	0.37	0.21	0.18	0.51
R&D	0	0.05	0.12	0.00	0.04
	1	0.02	0.06	0.00	0.01
ROA	0	-0.02	0.23	-0.01	0.07
	1	0.04	0.09	0.01	0.07
Tobin’s Q	0	1.86	1.55	1.05	2.06
	1	1.86	1.19	1.09	2.10
Sales Growth	0	1.56	22.94	0.97	1.18
	1	1.10	0.23	1.00	1.13
Financial Constraint	0	-0.27	0.81	-0.33	-0.18
	1	-0.38	0.12	-0.47	-0.29
Annual Exret	0	0.01	0.52	-0.25	0.18
	1	0.03	0.31	-0.16	0.19
Log(age)	0	2.11	0.62	1.79	2.56
	1	2.26	0.60	1.95	2.71
Risk Factor Length	0	7,191	5,288	3,668	9,299
	1	8,710	6,422	4,304	12,019
NIST Counts	0	95	71	48	122
	1	135	96	66	178

The table presents the summary statistics for key characteristics of both cyberattack victims and non-victims. There are in total 37,481 firm-year pairs, with 552 cyberattack victim cases, during the period 2006–2018. We show the mean, standard deviation, 25^{th} percentile, and 75^{th} percentile for each group, and present them side by side for comparison. Victims are denoted as $Victim = 1$, while zero indicates non-victims. The variables are defined in the Appendix. Risk Factor Length is the number of tokens in a firm’s 10-K Item 1A. NIST Term Counts is the number of times a firm mentions cybersecurity terms per NIST dictionary in its Item 1A in 10-K.

2.3.1 Constructing Predictors

Kamiya et al. (2021) document that a set of firm-specific variables are correlated with the probability that a firm experiences a cyberattack. We use these variables as the starting point in building our model. The variables include asset intangibility, the ratio of capital expenditures to total assets (CAPX/AT), the ratio of cash flows to total assets (CF/AT), firm size, net worth, R&D, ROA, Tobin's q, sales growth, financial constraint, prior year excess return over the CRSP value-weighted market return, and the natural log of firm age. In addition, we explore soft information about a firm's cyber vulnerability. One source is a firm's self-assessment of its cyber risk disclosed in the "Item 1A: Risk Factors" section in firms' annual 10-K filings. The SEC requires that firms disclose the most significant risk factors.[12] Per this requirement, firms started to include this section prominently and regularly from 2006. Firms are in a unique position to assess their own risk level given their private information and familiarity with the firm's operations. In addition, this risk disclosure is forward-looking, providing useful information about the perceived exposures. The nature and intensity of the discussion on cyber-related risk in their 10-K filings should be a useful indicator of their firms' vulnerability to cyberattacks.

To measure the intensity of perceived cyber risk, we count the number of times a firm mentions cyber risk-related terms in the risk factors section in its 10K filing as well as the length of the entire section. In this way, we can capture the information about a firm's self-assessment of both its overall risk exposure and exposure specific to cyber risk.

To identify cyber risk terms, we exploit a specialized dictionary compiled by the National Institute of Standards and Technology (NIST), named the "Glossary of Key Information Security Terms."[13] An important feature worth noting is that the NIST dictionary was compiled in 2006, suggesting that cyber risk was already a concern with governmental bodies. Secondly, it avoids a potential look-ahead bias since our sample starts in the year 2006.[14]

Next, we turn to firms' annual 10-K filings. Since the SEC mandate for disclosing risk factors

[12] See *SEC*, 2005, https://www.sec.gov/rules/final/33-8591.pdf.

[13] *NIST*, April 25, 2006, https://www.nist.gov/publications/glossary-key-information-security-terms.

[14] A contemporaneous study by Jamilov et al. (2020) uses a dictionary with a different set of words and phrases. We show in Section **??** that our results are robust to using alternative dictionaries.

came into effect in 2006, we start our sample in 2006. We obtain the filings from 2006 to 2018 through the University of Notre Dame Repository and extract their risk factor sections.[15] We use the central index key (CIK) to match the filing firms to Compustat firms. Then we count the number of tokens per risk factor text block, only keeping the blocks with at least 100 tokens.[16]

The final step in constructing the cyber risk disclosure variable is to measure the intensity with which cybersecurity-related terms are mentioned. We apply the following procedure: first, we measure the maximum number of tokens in each term in each dictionary. We find that the NIST dictionary contains terms that can be as long as seven tokens. Second, to control for the fact that there might be common boilerplate language that most firms use, we eliminate the common terms used by firms in each year as in Hoberg and Philips (2016). Third, we transform each risk factor text block into a collection of $Ngram$s, where $N \in [1, 2, 3, 4, 5, 6, 7]$. Each $Ngram$ is a collection of every possible combination of adjacent N tokens in each document.[17] Finally, we count how many times each cybersecurity term occurs in each $Ngram$-transformed text block.

We use the length of the risk factor section and the NIST term frequency as two separate variables for two reasons. First, the length of the risk factor section measures the number of tokens, or uni-grams, while the NIST term frequency is the number of times each term is mentioned. Here the term ranges from uni-grams to seven grams, and thus it is inappropriate to scale the frequency by the length of the section. Second, we keep both variables in the feature set to maximize the power of the machine-learning techniques.

2.3.2 Constructing the Predictive Model

2.3.2.1 In-sample Fit

Before building the predictive models, we examine the in-sample explanatory power of the regressors we select in-sample. In particular, we perform in-sample logit regressions, which

[15] We use an algorithm to match the section between the beginning of Section 1A and either Section 1B or Section 2.

[16] A token is an instance of a sequence of characters in some particular document that is grouped together as a useful semantic unit for processing (Schütze et al. (2008)). Here we refer to each term as a token after preprocessing the text, such as dropping stop words.

[17] See, e.g., Damashek (1995) for description of Ngrams.

include the accounting and text-based variables, which are cross-sectionally standardized to have means of zero and standard deviations of one. Following Petersen (2009), we cluster standard errors by both firm and year.

The results in Table **??** of the Appendix show that many firm characteristics have strong associations with the probability of future cyberattacks. In terms of magnitudes, firm size is particularly important, with larger firms more likely to be attacked. We also find that the text-based cyber risk count variables have positive predictive power for future cyberattacks. Interestingly, as shown in Column (1), the total number of tokens in the Risk Factors section of a firm's 10-K filing has a negative association with the probability of future cyberattacks. This result suggests that firms with more thorough discussions of their risk factors may also have better risk management practices.

2.3.2.2 Building Machine Learning Predicting Models

To build an effective cyberattack forecasting model, there are two main considerations. First, an in-sample fit could induce look-ahead bias and overfitting. Second, cyberattacks are rare events, which result in a highly imbalanced sample, thereby introducing biases into the Maximum Likelihood Estimator (King and Zeng (2001)).

To address the possible look-ahead bias, we follow a recursive and expanding prediction procedure. Namely, we start from the year 2006, run a predictive model, save the coefficients, and then use them to fit the data in 2007. We do this recursively, so the final step is to use the data from 2006 to 2017 to train the model to fit the year 2018. In this way, we avoid the look-ahead bias and generate true out-of-sample (OOS) predictions.

Next, to address overfitting and maximize OOS performance, we introduce machine learning classification models to improve the predictive power. In particular, we select logistic ridge regression as our main model.[18] The logistic ridge regression combines the logistic regression with ridge regularization, which uses an L-2 penalty.[19] The objective function for logistic ridge

[18] We show in Section **??** additional results based on Ensemble methods, and find our results to be robust to model choice.

[19] Another popular algorithm in asset pricing is LASSO, which uses an L-1 penalty to achieve model sparsity. Since

regression is:

$$\min_{\beta_0, \boldsymbol{\beta} \in \mathbb{R}^{p+1}} -\Big[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + \boldsymbol{x_i}^T \boldsymbol{\beta}) - \log(1 + e^{\beta_0 + \boldsymbol{x_i}^T \boldsymbol{\beta}})\Big] + \frac{1}{2}\lambda \|\boldsymbol{\beta}\|_2^2 \tag{2.1}$$

where p is the number of parameters, and λ controls the regularization strength.

To fully exploit the conditioning information and the logistic ridge regression's regularization capability, we perform the polynomial transformation of our regressors.[20] A polynomial transformation of degree d converts the variables to all possible interactions and powers up to degree d. In our implementation, we choose a polynomial transformation of degree 4.

In each training window, we tune hyperparameters for optimal performance via three-fold cross-validation, using the training data in each recursive sample period. Cross-validation is done through the following procedure: for the training data in each window, we randomly split the training data into three folds, i.e., three subsets of equal size; we use two of the three folds to fit the model, and one remaining fold as the "validation set" to find the best estimator; we iterate this procedure three times and find the best estimator, in terms of out-of-sample performance metric (based on the validation set of each iteration); then we use this estimator for prediction on the test data. Since regularization is sensitive to the unit of variables, we scale the variables before the training process (Hastie et al., 2017).

Since our sample is highly imbalanced, we use the stratified K-fold strategy of Zeng and Martinez (2000) to ensure that the class distribution is maintained across all folds. In addition to tuning the penalty factor λ, we also tune the class weight parameter in the loss function, using a heuristic proposed by King and Zeng (2001).[21]

we are primarily interested in the forecasting performance of the model rather than variable selection, we focus on logistic ridge regression. With LASSO, we are able to report weaker but consistent results.

[20] See Hastie et al. (2017) for an introduction of polynomial transformation. This function is implicitly used in kernel methods such as support vector machines.

[21] King and Zeng (2001) shows that in highly imbalanced learning problems, the coefficients can be biased towards the majority class, effectively rendering the model useless. They propose several measures to mitigate the bias, including the prior correction procedure and weighted logistic regression. We use the second option as it is less costly and shows similar performance compared with the prior correction procedure. The weighted logit equation is: $\ln \mathcal{L}_w(\beta|y) = w_1 \sum_{Y_i=1} \ln \pi_i + w_0 \sum_{Y_i=0} \ln(1-\pi_i)$. The intuition for the weighted logistic is that the objective function is weighted by a coefficient w that is inversely related to the sample class distribution. Consequently, the minority class gets weighted more heavily, thus mitigating the bias. We tune w via cross-validation to add extra flexibility.

2.3.2.3 Filtering the Sample

In our study, we wish to capture significant cyberattacks with meaningful impacts on the victim firms. In a contemporaneous study, Florackis et al. (2022) select events that are prominently featured in media reports as a filter that identifies important attacks. We take a similar but more direct approach: we use a filter based on the 7-day cumulative abnormal returns (CAR) around the events. Specifically, we first calculate the [-3,3] window CAR for each event using a four-factor model, the Fama-French three factors and the momentum factor, as the benchmark. Then we choose the events with an absolute value of CAR greater than 1%; we consider these events as important to investors based on the stronger market reaction.[22] After using this filter, our sample consists of 394 cyberattacks from 2007 to 2018.

2.3.2.4 Evaluating the Forecasting Performance

To evaluate the performance of the logistic ridge regression against the simple logistic regression, we follow the machine learning literature and choose three commonly used metrics of forecasting performance: the F1-score, G-mean, and Balanced Accuracy.[23]

The building blocks for these metrics include recall (also called sensitivity), precision, and specificity.

$$Recall = Sensitivity = \frac{True\,Positives}{True\,Positives + False\,Negatives}$$

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives}$$

$$Specificity = \frac{True\,Negatives}{True\,Negatives + False\,Positives}$$

The F1 score is the harmonic average of recall and precision. G-mean is the geometric average of recall and specificity. Balanced Accuracy is the arithmetic average of sensitivity (recall) and specificity.

[22] We choose 1% because Kamiya et al. (2021) document that attacked firms on average experienced -1% CAR around the events. Our results are robust to different specifications. We show in Section **??** that using 0.5% absolute CAR as the filter or using no filter produces similar results.

[23] See Brodersen et al. (2010); Yue et al. (2007); He and Garcia (2009) for discussions of these metrics.

Table 2.2 compares the average forecasting performance of the logistic ridge regression with that of the commonly used simple logistic regression over our sample period. Figure 2.3 presents the time series of the performance metrics.

Table 2.2 Performance Metrics

Metrics	F1-score: Harmonic Mean	G-mean: Geometric Mean	Balanced Accuracy
Logit	0.004	0.014	0.501
Logistic Ridge	0.063	0.599	0.650
Logistic Ridge - Logit	0.059***	0.584***	0.149***
	(0.010)	(0.043)	(0.022)
Note:			*p<0.1; **p<0.05; ***p<0.01

This table reports the time-series mean out-of-sample performance metrics across the recursive windows for logit and logistic ridge. F1-score is the harmonic mean between precision and recall. The geometric mean is the geometric mean between sensitivity and specificity. Balanced accuracy is the arithmetic mean between sensitivity and specificity. Each recursive window contains $n + 1$ years of data, with $n \in [1, 10]$ as the training data and +1 the subsequent year as the test data. There are in total 11 windows, starting at the year 2007 as the training data and 2008 as the test data. The final window consists of the years 2007 - 2017 as the training data and 2018 as the test data. The first two rows are the means of each corresponding metric for 11 windows. The third row represents the mean difference between the logistic ridge and logit, and the fourth row computes the standard error of the difference in means.

The results show that the logistic ridge regression strongly outperforms the baseline logistic regression across all three performance metrics. For instance, the average F1 score of the logistic ridge regression is approximately 15 times as large as that of the logistic regression, and the average G-mean of the logistic ridge regression is approximately 40 times as large as that of the logistic regression. The Balanced Accuracy of the logistic ridge regression exceeds that of the logistic regression by approximately 30%. The strong outperformance of the logistic ridge regression is persistent through time.

To understand the drivers of the superior performance of the logistic ridge regression, we plot the aggregate (or equivalently, mean) confusion matrices for the logistic ridge and logistic regressions

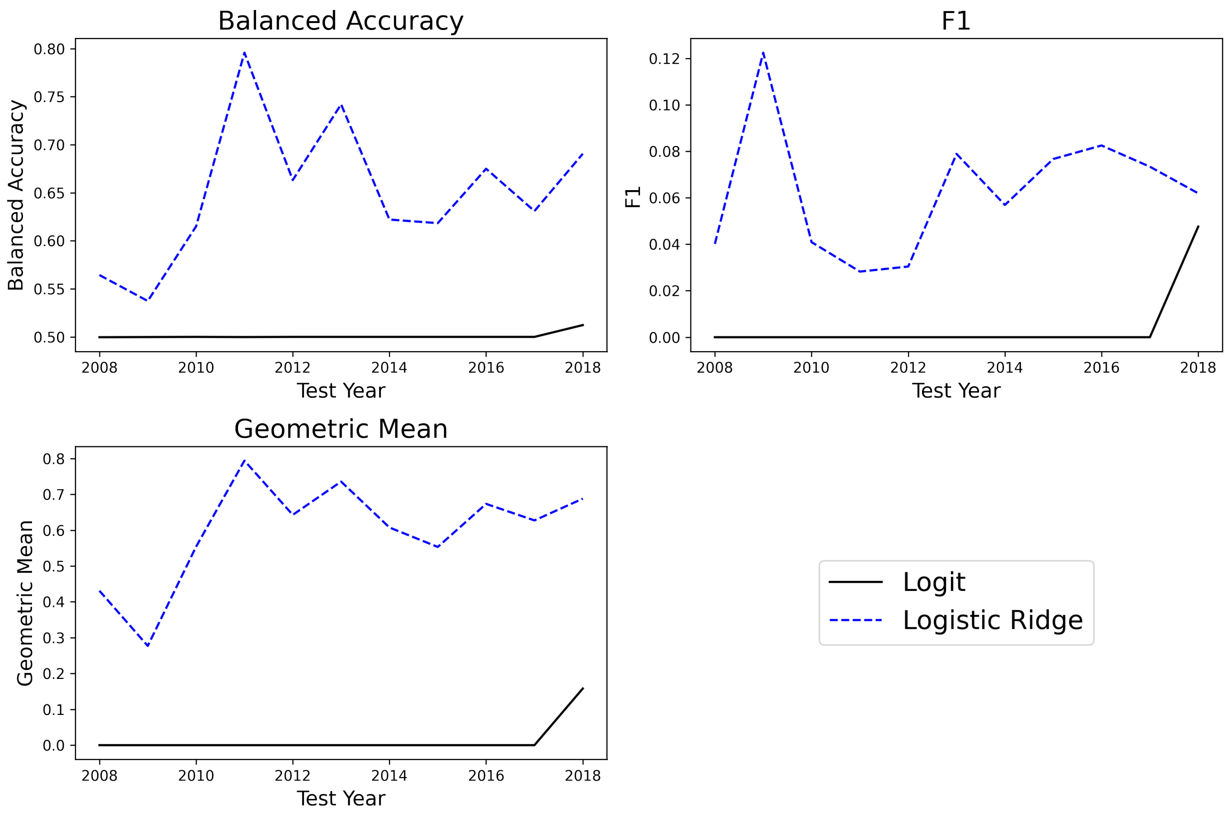


Figure 2.3 Out-of-Sample Performance Metrics

The figure depicts the time series of three out-of-sample performance metrics for both logit and logistic ridge regression across the test years. The three metrics are F1 score, G-mean, and Balanced Accuracy. F1-score is the harmonic mean between precision and recall. The geometric mean is the geometric mean between sensitivity and specificity. Balanced accuracy is the arithmetic mean between sensitivity and specificity. The black solid line indicates logit, while the blue dashed line represents logistic ridge. The metrics are all measured against test samples, and thus the figure runs from 2008 to 2018.

in Figure 2.4. The rows of a confusion matrix are the true classes, while the columns are predicted classes. Each of the four quadrants indicates the total number of observations across the sample period classified per the prediction model.

In a confusion matrix, a superior forecasting technique tends to register higher numbers in the diagonal elements, which implies that the classifier more correctly classifies observations. The results show the source of the weakness of the simple logistic regression: it overwhelmingly predicts firms as negative cases and under-predicts the positive cases. Among the firms that experience a cyberattack the following year, the algorithm falsely predicts more than 99% of them as negative

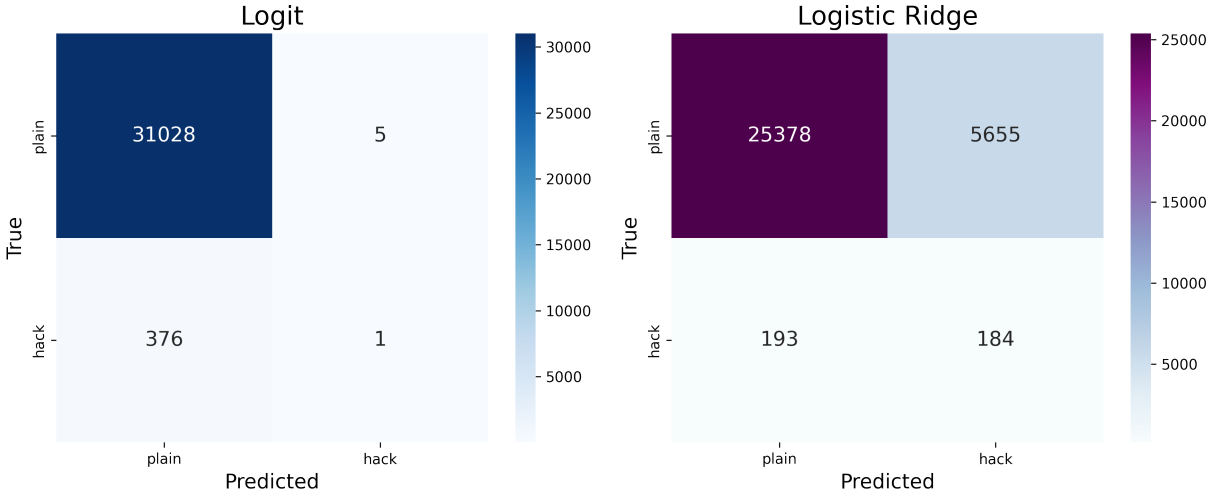


Figure 2.4 Confusion Matrices

The figure shows the aggregate/mean confusion matrices for logit and logistic ridge regression. That is, we aggregate the numbers in the four quadrants across the 11 test windows for each model. The rows are true classes, while the columns are predicted classes. Each quadrant presents the number of observations that the classification model allocates. The diagonal numbers represent the total number of observations that are correctly allocated to their respective classes. For example, for logit in the left panel, it correctly predicts 1 cyberattack case, while misallocating 376 cases as non-victims. On the other hand, logistic ridge regression in the right panel correctly predicts 184 cyberattack cases, while misallocating 193 cases in the non-victim class.

cases. In comparison, the logistic ridge regression falsely predicts approximately 50% of them as negative cases. In other words, the logistic ridge regression benefits from a jump in recall in exchange for a small decrease in precision, and thus achieves a better balance in performance.

In the rest of the paper, we use the cyber risk measure based on the logistic ridge regression as the primary measure. We present the results based on alternative machine learning techniques in Section **??** as robustness tests.

2.3.2.5 Characterizing the Firms with High Cyber Risk

Figure 2.5 shows the correlation coefficients between our cyber risk measure and firm characteristics. The left column presents the time-series averages of the cross-sectional Spearman correlations, and the heat map is based on the absolute value of the average correlation coefficient.

The results of average correlation coefficients indicate that firms with higher cyber risk tend to be larger, less financially constrained, and more profitable. The heat map also shows that these three

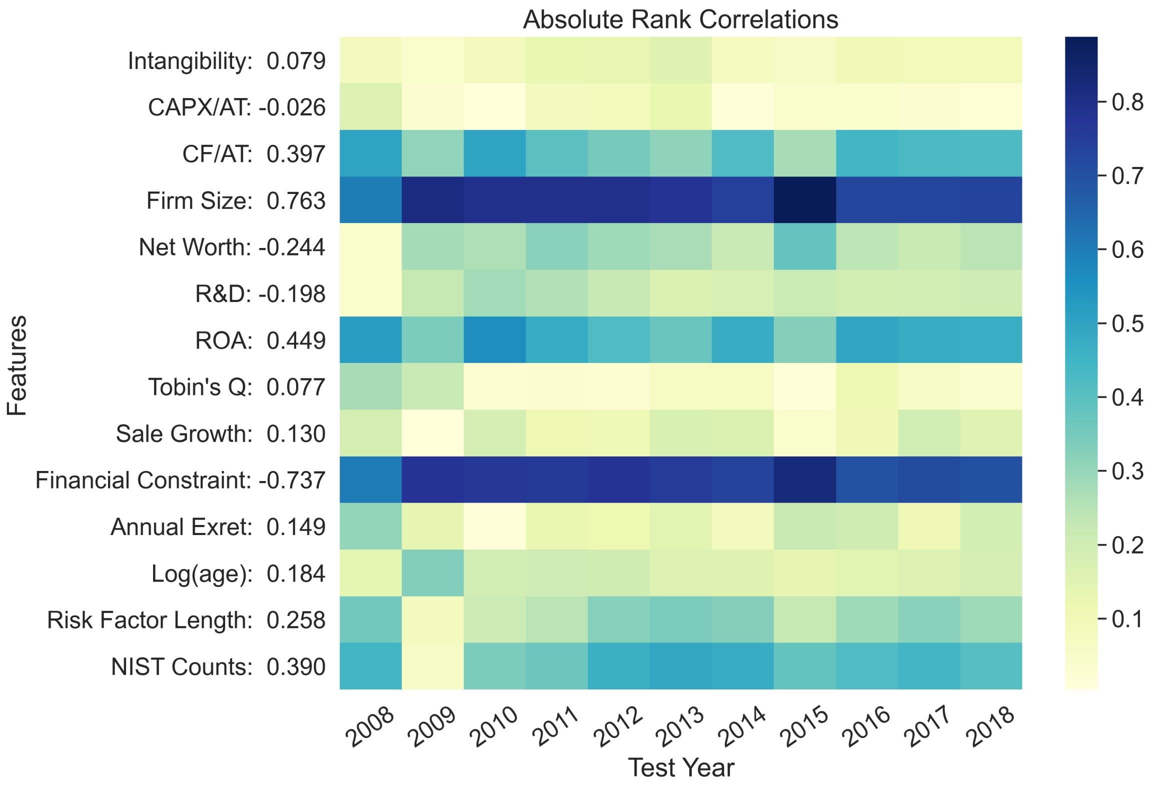


Figure 2.5 Absolute Rank Correlations between Predictors and Cyber Risk

The figure depicts the absolute value of rank correlations between each of the 14 predictor variables and the estimated cyber risk using logistic ridge regression for each test year, as well as the time-series mean of rank correlations for each variable. Specifically, for each test year, we compute the rank correlation between each variable and cyber risk measure, and then take the time-series mean for each variable and put them in the left column of the figure. Then we compute the absolute values of all the rank correlations of all variables across the test years. In so doing, we have 11 sets of coefficients, since the test years run from 2008 to 2018. We plot them in a heat map, as shown in the figure. The color depth represents how large the coefficient is relative to other variables. For example, firm size is consistently important across years, while CAPX/AT has been consistently less important.

characteristics tend to have high and stable correlations with cyber risk across time. In addition, the text-based variables have high correlations with cyber risk.

Before leaving this section, we examine the relation between our cyber risk measure and the measure of $Cyber\,Cosine$ developed by Florackis et al. (2022). Although both measures intend to capture firm-level cyber risk, the methodologies are distinct. It is therefore of interest to quantitatively examine the relation. We find that the correlation between the two measures is approximately 13.8%, which is indeed positive but moderate. Because the cyber cosine measure is purely text-based, it also makes sense to compare the correlation between their measure and the textual variables we use in our cyber risk forecasting model. We find that the correlation between cyber cosine and the "NIST Counts" is 0.40 and that between cyber cosine and "Risk Factor Length" 0.27, suggesting that our textual variables have a stronger comovement with $Cyber\ Cosine$. Overall, these results show that our cyber risk measure and the cyber cosine measure share a moderate amount of common information. In other words, both measures should be of value for studies on cyber risk.

2.4 Cyber Risk and Stock Returns

In this section, we study the relationship between cyber risk and stock returns. We start with cross-sectional analyses using both the Fama and MacBeth (1973b) regressions and portfolio analyses. Then we examine the time variation in the return spread between firms with high and low cyber risk, focusing on its comovement with other measures of aggregate cyber risk concern.

2.4.1 Cross-sectional Regressions

To examine whether firms with higher cyber risk compensate investors with higher average returns, we use Fama-MacBeth cross-sectional regressions to examine the incremental predictive power of the cyber risk measure for stock returns, controlling for well-known stock characteristics proposed in the previous literature. Specifically, at the end of each month from July of year t to June of year $t + 1$, we regress individual stock returns on our cyber risk measure estimated in year t and a set of firm characteristics. The cyber risk estimate is based on firms' accounting information for the fiscal year ending anytime in the calendar year $t - 1$, the cyberattack forecasting model parameters estimated using the accounting information available up to June in year $t - 2$, and the cyberattack incidents available up to the end of June of year $t - 1$. The year t ranges from 2008 to 2018. All regressors are cross-sectionally standardized to have a mean of zero and a standard deviation of one. Then we test whether the time-series averages of the regression coefficients are statistically significant. The standard errors are based on the Newey and West (1987) adjustment with 6 lags. For a stock to be included in the analysis, it is required to have a CRSP share code 10 or 11 and to have a closing price above \$5 at the end of June in year t.

Table 2.3 shows the results. In Column (1), we include our cyber risk measure, together with control variables including natural log of market cap, natural log of book-to-market ratio, gross profitability, asset growth, and momentum (measured as prior twelve to one-month return). Column (2) replaces our cyber risk measure with $Cyber\,Cosine$ of Florackis et al. (2022). Column (3) jointly includes our cyber risk measure and the cyber cosine measure. Column (4) further includes the past one-month return to control for short-term return reversal, idiosyncratic volatility (Ang et al., 2006), and illiquidity (Amihud, 2002). Column (5) adds organizational capital (Eisfeldt and Papanikolaou, 2013), CAPM beta, tail risk beta (Kelly and Jiang, 2014), coskewness (Harvey and Siddique, 2000), and net operating assets (Hirshleifer et al., 2004).

The results in Table 2.3 show a strong relation between our cyber risk measure and future stock returns. For instance, Column (1) indicates that a one-standard-deviation increase in cyber risk is associated with an approximately 20 basis-point increase in returns per month, or 2.4% per year. The effect is statistically significant at the 1% level. Consistent with Florackis et al. (2022), we find in Column (2) that the $Cyber\,Cosine$ measure has strong predictive power for future stock returns. When we include both variables in the same regression in Column (3), both our cyber risk measure and the cyber cosine measure have strong relations with future stock returns. In terms of magnitudes, the coefficient for our cyber risk measure is 0.186 and that for the cyber cosine measure 0.093. This result indicates that our cyber risk measure captures independent information about firms' cyber risk exposure as compared to the $Cyber\,Cosine$ measure of Florackis et al. (2022). Columns (4) and (5) further establish the robustness of the results when we include more control

Table 2.3 Fama-MacBeth Cross-sectional Regressions

	Dependent Variable: Returns				
	(1)	(2)	(3)	(4)	(5)
Cyber Risk	0.197***		0.186***	0.148**	0.212***
	(0.067)		(0.064)	(0.063)	(0.065)
Cyber Cosine		0.132***	0.093**	0.100**	0.049
		(0.049)	(0.043)	(0.043)	(0.044)
Log(MktCap)	-0.127	-0.006	-0.133	-0.159	-0.219*
	(0.158)	(0.139)	(0.152)	(0.128)	(0.120)
B/M	0.163	0.204	0.160	0.087	0.085
	(0.148)	(0.143)	(0.149)	(0.150)	(0.144)
GP	0.098	0.155	0.103	0.081	0.393**
	(0.108)	(0.112)	(0.107)	(0.100)	(0.152)
ATG	-0.247***	-0.225***	-0.249***	-0.244***	-0.112
	(0.055)	(0.058)	(0.055)	(0.056)	(0.076)
MOM	-0.089	-0.081	-0.091	-0.119	-0.202
	(0.181)	(0.183)	(0.182)	(0.188)	(0.179)
ST_Rev				-0.345***	-0.381***
				(0.107)	(0.126)
IdioRisk				-0.226	-0.101
				(0.146)	(0.191)
Illiquidity				0.134	0.258*
				(0.111)	(0.139)
Other Controls	NO	NO	NO	NO	YES
Observations	289,696	310,243	275,301	275,301	132,924
R-squared	0.031	0.030	0.032	0.046	0.078

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

This table reports Fama-MacBeth cross-sectional regressions of returns on firm characteristics and anomaly variables. For each month, we run cross-sectional regression of stock returns on firm characteristics, and then we average the coefficients across time series. The sample runs from July 2008 to June 2019. Column (1) includes the natural log of market cap, natural log of book-to-market ratio, gross profitability, asset growth, and momentum. Column (2) replaces our cyber risk measure with $Cyber\,Cosine$ (Florackis et al., 2022). Column (3) adds back our cyber risk measure. Column (4) further includes the past one-month return to control for the short-term return reversal, idiosyncratic volatility (Ang et al., 2006), and illiquidity (Amihud, 2002). Column (5) adds organizational capital (Eisfeldt and Papanikolaou, 2013), past market beta, tail beta (Kelly and Jiang, 2014), coskewness (Harvey and Siddique, 2000), and net operating assets (Hirshleifer et al., 2004). Standard errors are Newey-West standard errors with 6 lags and are reported in parentheses. Variable definitions can be found in Appendix A. Returns are measured as percentages, and all independent variables are cross-sectionally winsorized at [1%,99%] level and standardized to have zero mean and standard deviation of unity.

variables.

2.4.2 Portfolio Sorts

We also use the portfolio approach to examine the relationship between cyber risk and stock returns. At the end of each June from 2008 to 2018, we sort stocks into three tercile portfolios, four quartile portfolios, and five quintile portfolios according to the cyber risk measure. These portfolios are held until the end of June in year $t + 1$ and then rebalanced based on the updated cyber risk estimate when new information becomes available.

We create zero-cost hedge portfolios that buy high cyber-risk stocks and short low cyber-risk stocks. We compute the monthly equal- and value-weighted alphas on the hedge portfolios using a number of asset pricing models: the CAPM, Fama-French three-factor model (FF3), FF3 augmented by a momentum factor (FF4), Fama-French five-factor model (FF5), and FF5 augmented with a momentum factor (FF6).

Table 2.4 shows that the hedge portfolios generate large alphas across the different portfolio constructions and against the various benchmark models. The monthly alpha ranges from 40 to 80 basis points per month and is always statistically significant. These results show that investors in high cyber-risk stocks earn high average returns, which cannot be explained by these asset pricing models.

2.4.3 Comovement with the Index of Cybersecurity

In addition to the cross-sectional variation in average returns, we examine the time variation in the relative returns between stocks with high and low cyber risk. In this subsection, we relate the return spread to the index based on an independent survey conducted by New York University on the perception of cyber risk among industry experts: the Index of Cybersecurity (ICS).[24] Each month, the survey queries experts such as chief risk officers, chief information security officers, selected academicians engaged in fieldwork, and selected security product vendors' chief scientists about their perceived level of cyber risk, from which an aggregate measure of cyber risk is built.

[24] ICS, *NYU Engineering*, https://wp.nyu.edu/awm1/. According to the website, "The Index of Cyber Security is a measure of perceived risk. A higher index value indicates a perception of increasing risk, while a lower index value indicates the opposite."

Table 2.4 Portfolio Analyses

	Value-Weighted			Equal-Weighted		
model	Tercile	Quartile	Quintile	Tercile	Quartile	Quintile
CAPM	0.574***	0.765***	0.843***	0.508***	0.645***	0.713***
	(0.176)	(0.207)	(0.227)	(0.154)	(0.171)	(0.180)
FF3	0.458***	0.658***	0.739***	0.437***	0.581***	0.657***
	(0.122)	(0.190)	(0.219)	(0.131)	(0.154)	(0.157)
FF4	0.461***	0.660***	0.742***	0.433***	0.577***	0.653***
	(0.127)	(0.193)	(0.226)	(0.121)	(0.141)	(0.150)
FF5	0.431***	0.536***	0.577***	0.425***	0.541***	0.616***
	(0.116)	(0.157)	(0.179)	(0.138)	(0.160)	(0.165)
FF6	0.438***	0.540***	0.585***	0.419***	0.533***	0.611***
	(0.124)	(0.161)	(0.190)	(0.131)	(0.151)	(0.163)

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

This table provides various portfolio analyses for the universe of stocks sorted on predicted ex-ante cyberattack probabilities. The prediction model used here is the logistic ridge regression. At the end of each June from 2008 to 2018, we sort stocks into three tercile portfolios, four quartile portfolios, and five quintile portfolios according to the estimated cyber risk measure. These portfolios are held until the end of June in year $t+1$ and then rebalanced based on the updated cyber risk estimate when new information is available. We create zero-cost hedge portfolios that buy high cyber-risk stocks and short low cyber-risk stocks. We compute the monthly equal- and value-weighted alpha on the hedge portfolios using a number of asset pricing models: the CAPM, Fama-French three-factor model (FF3), FF3 augmented by a momentum factor (FF4), Fama-French five-factor model (FF5), and FF5 augmented with a momentum factor (FF6). The left half panel reports value-weighted results, while the right half panel reports equal-weighted results. Standard errors are Newey-West standard errors with 6 lags and are included in parentheses.

To examine the relationship between our annual firm-level cyber risk measure and the monthly aggregate ICS, we project both variables onto the return space. For each stock in our sample, we estimate its return sensitivity to the monthly percentage change in the ICS. That is, we perform bivariate regressions of monthly excess returns of individual stocks on the monthly changes in the ICS and the excess returns on the market portfolio. Stocks with a high (low) ICS beta tend to have lower (higher) cyber risk exposures based on the ICS. Therefore, a long-short portfolio that buys stocks with a high ICS beta and sells those with a low ICS beta should deliver higher performance when aggregate cyber risk concern is high. If our firm-level cyber risk measure and the ICS capture some common component of cyber risk in the economy, we would expect the portfolio that buys high cyber-risk stocks and sells low cyber-risk stocks to have a negative correlation with the ICS factor-mimicking portfolio.

We apply a 12-month rolling window to estimate the ICS beta for individual stocks, requiring at least six observations for the statistical estimation. We form five quintile portfolios at the end of each month from December 2015. Then we repeat this process for each subsequent month. Therefore, we have monthly ICS beta portfolios from December 2015 to June 2019, with concurrent observations for our cyber risk-based quintile portfolios.

Panel A of Figure 2.6 shows the performance of the portfolio that buys stocks with high cyber risk and sells those with low cyber risk (the solid black line) against the return on the portfolio that buys high ICS beta stocks and sells low ICS beta stocks (the dotted red line). Consistent with our conjecture, the two series show a strong negative comovement. The time-series correlation coefficient is -41.2% and statistically significant.

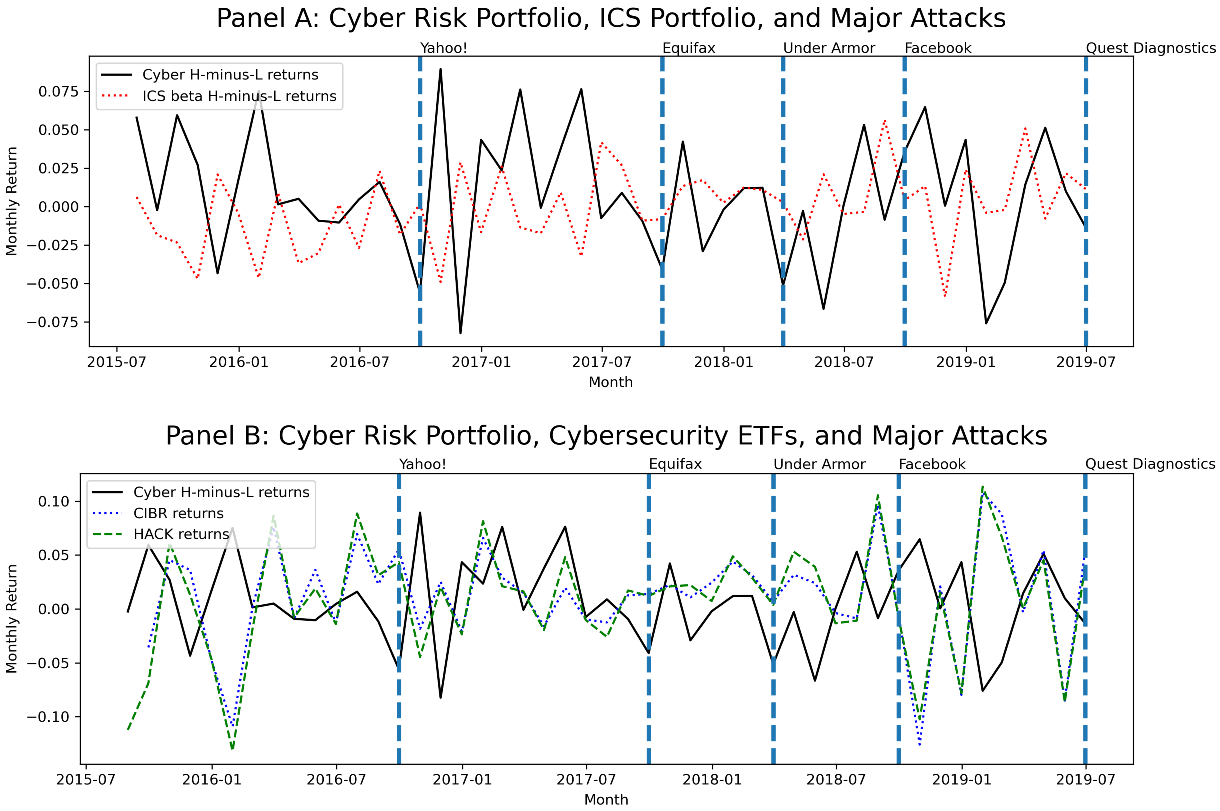
2.4.4 Comovement with the Cybersecurity ETFs

In this subsection, we use the return on cybersecurity index ETFs as another proxy for aggregate concern about cyber risk. The conjecture is that when investors have stronger concerns about cyber risk, cybersecurity ETFs tend to outperform. Our analyses include two such ETFs: the First Trust Nasdaq Cybersecurity ETF (CIBR)[25] and ETFMG Prime Cyber Security ETF (HACK).[26] Both ETFs track a set of firms that provide cybersecurity services. Table 2.5 shows their top ten holdings.

Panel B of Figure 2.6 plots the monthly returns on the spread portfolio that buys high cyber risk stocks and sells low cyber risk stocks (the solid dark line) against the performance of the cybersecurity ETFs (the dotted blue and dashed green lines). It shows a strong negative comovement. The time-series correlations are below -40% for both ETFs. If we perform a time series regression of our cyber risk portfolio return against the two ETF returns individually, the coefficients are -0.405 and -0.312, respectively, for CIBR and HACK, and are statistically significant at the 1% level. This result provides further validation for our cyber risk measure.

[25] CIBR, https://www.ftportfolios.com/retail/etf/etfsummary.aspx?Ticker=CIBR.

[26] HACK, https://etfmg.com/funds/hack/.



Panel A: Cyber Risk Portfolio, ICS Portfolio, and Major Attacks

Panel B: Cyber Risk Portfolio, Cybersecurity ETFs, and Major Attacks

Figure 2.6 Comovement with ICS and Cybersecurity ETFs & Major Cyberattacks

The figure shows the comovement of the return series of our high-minus-low cyber risk portfolio and that of the ICS beta portfolio, and that of two cybersecurity ETFs. In Panel A, we plot the return series of our high-minus-low cyber risk portfolio and ICS beta portfolio. The ICS beta portfolio is constructed at a monthly frequency by sorting stocks into quintile portfolios based on stocks' return betas with respect to the monthly change of the ICS index, controlling for market excess returns during the past 12 months. The two series have a correlation of -41.2%. In Panel B, we plot the monthly returns of the First Trust Nasdaq Cybersecurity ETF (CIBR) and ETFMG Prime Cyber Security ETF (HACK) from August 2015 to June 2019. The time-series correlation between the returns on the long-short cyber risk portfolio and the ETFs are -47.6% and -41.5% for CIBR and HACK, respectively. In both panels, we also mark the months that major cyberattacks happened, where the major attacks are according to a Bloomberg report (URL: https://www.bloomberg.com/graphics/corporate-hacks-cyber-attacks/?sref=GlJWBQ7Q). We screen the major attacks to only include public firm attacks during our sample period, which include Yahoo!, Equifax, Under Armor, Facebook, and Quest Diagnostics. Returns and events are plotted at month ends.

Table 2.5 Cybersecurity ETFs: Top 10 Holdings

CIBR		HACK	
Holding	Weight	Holding	Weight
CrowdStrike Holdings, Inc. (Class A)	7.49%	Blackberry Ltd.	3.57%
Zscaler, Inc.	7.08%	Cisco Sys Inc.	3.13%
Cisco Systems, Inc.	5.60%	Palo Alto Networks Inc.	2.81%
Accenture Plc	5.25%	Cyberark Software Ltd.	2.80%
Splunk Inc.	4.41%	Ping Identity Hldg Corp	2.77%
FireEye, Inc.	3.41%	FireEye Inc.	2.76%
Palo Alto Networks, Inc.	3.38%	Sumo Logic Inc.	2.71%
Proofpoint, Inc.	3.28%	Commvault Systems Inc.	2.69%
SailPoint Technologies Holdings, Inc.	3.27%	Cloudflare Inc.	2.64%
Fortinet, Inc.	3.22%	Qualys Inc.	2.64%

This table presents the top ten holdings of two cybersecurity ETFs, respectively: First Trust Nasdaq Cybersecurity ETF (CIBR) and ETFMG Prime Cyber Security ETF (HACK). The holdings are as of February 2021.

As a still further validation, we hypothesize that when major cyberattacks happen, high cyber-risk firms would underperform low cyber-risk firms. To examine this conjecture, we first identify major cyberattacks via a Bloomberg report.[27] Because the list consists of private and public firms, we further screen the list to include only public firms to match our sample. We then manually check the dates to ensure they reflect when the attacks became public. The final list includes the following firms: Yahoo! (September 2016), Equifax (September 2017), Under Armor (March 2018), Facebook (September 2018), and Quest Diagnostics (June 2019). We mark these events in both panels of Figure 2.6. The figure shows clearly that our cyber risk high-minus-low portfolio underperformed when major cyberattacks happened, while the ICS beta portfolio and cybersecurity ETFs overperformed, consistent with our conjecture. These results provide further evidence for the validity of our cyber risk measure.

[27] Bloomberg report: https://www.bloomberg.com/graphics/corporate-hacks-cyber-attacks/?sref=GlJWBQ7Q.

2.5 Is Cyber Risk Priced? Further Identification

The preceding results show a strong relation between our cyber risk measure and future stock returns, which is consistent with the view that cyber risk is priced in the stock market. In this section, we provide further tests that strengthen the identification of the cyber risk premium.

2.5.1 Industry Competition

We start by exploiting the variation along the dimension of industry competition. Kamiya et al. (2021) show that when a firm is hacked, its peer firms in the same industry tend to experience stock price drops around the announcement. This result is consistent with the view that increased awareness of cyber risk exposure for peer firms leads the stock market to reprice their stocks. Building on this observation, we hypothesize that the average market reaction to peer firms is confounded by the nature of product market competition in the industry. For instance, if the hacked firm is a weaker player in the product market, the incidence of a cyber attack is likely to provide its stronger competitor an opportunity to increase market share. Thus, the stock market reaction to the cyber risk exposure of the peer firm can be muted by the stock market's perception of its improved product market opportunities. However, this confounding effect is likely to be weaker when the hacked firm is a stronger player in the industry: It is harder for a weaker competitor to exploit the potential opportunity to the same extent.

To empirically test this hypothesis, for each hacked firm, we identify its peers in the same industry based on the Fama-French 48-industry classification (Fama and French, 1997). We construct a variable "Strong Victim" to capture the relative competitive position of the hacked firm (victim) to a peer firm: It equals one if the victim has a higher market share (firm sales over industry sales) than that of the peer firm, and zero otherwise. The idea is that if the victim is a stronger (weaker) competitor of the peer firm, the peer firm is less (more) likely to increase its market share in response to a cyberattack on the victim firm. As a result, the negative market reaction to the peer firm around the time period when the data breach of the victim becomes public would be stronger. This is because it is a cleaner test of the market price reaction of a peer's stock as investors re-calibrate their estimate of its cyber risk in response to the victim getting hacked. That is, in a regression of the [-3,+3] window cumulative abnormal return (CAR) of the peer firm on the [-3,+3] window CAR of the victim firm around the date when the data breach is announced, the coefficient for the interaction of the “Strong Victim” dummy variable and the victim’s CAR is expected to be positive.

Table 2.6 presents the results based on the cyberattack events when the victim firm experiences a CAR of lower than -1% during the [-3,+3] window. In Column (1), we confirm the analysis in Kamiya et al. (2021) that when focal firms experience a cyberattack, peer firms in the same industry also suffer, demonstrated by the statistically significant and positive coefficient for “Victim CAR”. Column (2) shows that the “Strong Victim” variable itself has no effect on the stock market reaction of peer firms. However, when we interact “Strong Victim” with “Victim CAR”, the coefficient for the interaction is statistically significant and positive. The spillover of the negative stock market reaction from the victim to its peer firms is stronger when the confounding effect from the product market is weaker. This reinforces our claim that cyber risk is an important driver of stock prices.

2.5.2 Product Similarity as a Proxy for Data Similarity

We note that firms with valuable data holdings tend to be particularly vulnerable to cyberattacks. Based on this observation, we seek to identify firms with data holdings similar to the victim firm. Hoberg and Philips (2016) propose an interesting measure that captures the similarity of products and services between two firms. If two firms provide similar products and services to their customers, it is likely that the data generated and stored by the two firms would be similar. Based on the Hoberg and Philips (2016) product similarity score, we identify firms providing similar products and services to those of the victim firm. Since these firms should have a higher probability of experiencing a cyberattack due to their data similarity to the hacked firm, a public release of information about a hack would lead the stock market to update their beliefs about the cyber risk exposure of these firms, resulting in a larger stock price drop for them.

To test this hypothesis, we use the event study framework similar to the preceding tests, focusing on the interaction between “Data Similarity" and “Victim CAR". In Table 2.7, we find that firms providing products and services similar to those of the victim firm experience a more negative

Table 2.6 Stock Price Reaction of Peer Firms to Cyberattacks and Industry Competition

	Dependent Variable: Peer CAR [-3,3]		
	(1)	(2)	(3)
Victim CAR	0.046***	0.046***	0.051***
	(0.015)	(0.015)	(0.013)
Strong Victim		-0.000	0.001
		(0.002)	(0.002)
Victim CAR × Strong Victim			0.025***
			(0.008)
Size	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
B/M	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
Constant	0.002	0.002	0.002
	(0.005)	(0.006)	(0.006)
Observations	57,194	57,194	57,194
R-squared	0.008	0.008	0.008
Industry & Year FE	YES	YES	YES
Industry & Event Cluster	YES	YES	YES

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

This table examines how industry competition influences the stock market reaction of peer firms to cyberattacks on victims. Peer firms are defined as those in the same Fama-French 48 industry (Fama and French, 1997) as the victim firms that experienced a cyberattack. We select only those cyberattack incidents when victims experienced lower than -1% CAR around the [-3,+3] window, consistent with the screening criteria used in our main results. The CAR is defined as the cumulative daily abnormal returns based on the Fama-French three-factor model augmented by the momentum factor, with an estimation window of 100 days. We require a minimum number of valid returns of 70 days in the estimation window, and a gap of 50 days between the end of the estimation window and the beginning of the event window. Victim CAR is the [-3,+3] window cumulative abnormal return of the focal firm around cyberattack events. Peer CAR is the [-3,+3] window cumulative abnormal return of the peer firm around cyberattack events. "Strong Victim" is a dummy variable if the peer firm's sales are lower than the victim firm in the previous year or zero otherwise. Industry and year-fixed effects are included. Standard errors are clustered at the industry and event levels.

market reaction when the data breach of the victim becomes public, which supports our hypothesis. In terms of magnitudes, because the data similarity measure has a mean of 0.022 and a standard deviation of 0.045, a one-standard-deviation increase in data similarity is associated with an increase in the response of peer firm's stock price to the CAR of the victim firm by 0.037 (0.045 × 0.82).

Table 2.7 Stock Price Reaction of Firms with Data Similarity to Cyberattacks

	Dependent Variable: Peer CAR [-3,3]		
	(1)	(2)	(3)
Victim CAR	0.046***	0.049***	0.038**
	(0.015)	(0.017)	(0.016)
Data Similarity		-0.055***	-0.011
		(0.010)	(0.009)
Victim CAR × Data Similarity			0.820***
			(0.237)
Size	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
B/M	-0.001	-0.000	-0.001
	(0.001)	(0.001)	(0.001)
Constant	0.002	0.003	0.003
	(0.005)	(0.005)	(0.005)
Observations	57,191	57,191	57,191
R-squared	0.008	0.008	0.009
Industry & Year FE	YES	YES	YES
Industry & Event Cluster	YES	YES	YES
Note:			*p<0.1; **p<0.05; ***p<0.01

This table uses product similarity as a proxy for data similarity to study the impact of cyberattacks on firms with similar data holdings. First, we identify firms in the same Fama-French 48 industry (Fama and French, 1997) as the victim firms that experienced a cyberattack. Then we use the product similarity score from Hoberg and Philips (2016) as our proxy for data similarity between two firms. Other variables are as defined in Table 2.6. Industry and year-fixed effects are included. Standard errors are clustered at the industry and event levels.

Building on the success of the Hoberg and Philips (2016) product similarity measure to capture data similarity, we construct a firm-level Composite Data Similarity measure that averages a firm's data similarity with all the victim firms experiencing a cyberattack in the previous year:

$$CompositeDataSim_{i,t} = \frac{1}{m} \sum_{j=1}^{m} DataSim_{i,j,t}, \tag{2.2}$$

where i represents individual firms in the Compustat universe, j indexes firms that have been attacked in year t, and $DataSim$ is the product similarity score measure in Hoberg and Philips (2016). Thus $CompositeDataSim$ is the average data similarity of each firm to the group of attacked firms in the previous year.

Since a firm holding valuable data similar to previously attacked firms is a more likely target, we expect its composite data similarity score to be a positive contributor to a firm's cyber risk. To examine this conjecture, we regress both our cyber risk measure and the cyber cosine measure proposed by Florackis et al. (2022) on the composite data similarity. To make the coefficients comparable, we standardize each of the continuous variables to be have a mean of zero and a standard deviation of unity. The results in Table 2.8 are consistent with our hypothesis. The coefficients for the "Composite Data Sim" are positive and statistically significant across the different specifications. A one-standard-deviation increase in "Composite Data Sim" is associated with a 0.012 standard deviation increase in our "Cyber Risk" measure, and a 0.038 standard deviation increase in the "Cyber Cosine" measure, when we control for the effects of firm size and book-to-market ratio. These results support the idea that valuable data holdings tend to attract the attention of hackers, which increases the cyber risk of a firm.

2.6 Conclusion

In this paper, we use machine learning algorithms to develop an ex-ante cyber risk measure for individual firms, which has a superior ability to forecast the occurrence of future cyberattacks. We find that firms with higher cyber risk, according to this measure, earn higher average stock returns, which cannot be explained by standard asset pricing models. In times when these firms underperform, cybersecurity experts tend to have higher concerns about cyber risk, and cybersecurity exchange-traded funds outperform. Further evidence based on product market competition and data similarity between firms provides further support to the notion that cyber risk is an important determinant of expected returns in increasingly digitized economies.

Table 2.8 Cyber Risk and Data Similarity

	(1)	(2)	(3)	(4)
VARIABLES	Cyber Risk		Cyber Cosine	
Composite Data Sim	0.028***	0.012***	0.061***	0.038***
	(0.007)	(0.004)	(0.016)	(0.010)
Log(MktCap)		0.137***		0.200***
		(0.003)		(0.013)
B/M		0.021***		-0.013
		(0.002)		(0.011)
Constant	0.000***	-0.013***	0.042***	0.024***
	(0.000)	(0.000)	(0.000)	(0.001)
Observations	31,106	31,084	29,797	29,776
R-squared	0.872	0.887	0.470	0.506
Industry FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Note:			*p<0.1; **p<0.05; ***p<0.01	

This table presents the results of regressing our cyber risk measure and the cyber cosine measure from Florackis et al. (2022) on the composite data similarity measure in the previous year. We construct the "Composite Data Sim" variable as follows:

$$CompositeDataSim_{i,t} = \frac{1}{m}\sum_{j=1}^{m} DataSim_{i,j,t},$$

where i represents firms in the Compustat universe, j indexes firms that have been attacked in year t, and $DataSim$ is the product similarity score measure in Hoberg and Philips (2016). In Columns (2) and (4), we control for size and book-to-market ratio. Every continuous variable is standardized to have a mean of zero and a standard deviation of one. We include industry (Fama and French 48-industry classification) fixed effects and year fixed effects. Standard errors are clustered at the industry level.

Our study suggests interesting avenues for future research. First, we have followed a bottom-up approach to estimate the cyber risk premium, starting with firm-level estimates of cyber risk. Another approach, which is top-down, can examine the systemic impact of a large attack on major players in the economy. For instance, Eisenbach et al. (2020) models how cyber attacks on large banks can influence the financial system and even the economy through network effects. It would be of interest to explore the connections between these two broad approaches.

Second, a limitation of our study is that we focus on the likelihood of cyberattacks as a proxy for cyber risk but do not study the scale of the loss resulting from cyberattacks. Although the probability of cyberattacks is likely to be correlated with the severity of cyberattacks, these two dimensions of cyber risk can contain different information. Our research makes a useful first step toward understanding the implications of cyber risk for asset prices. Future research can benefit from investigating these two dimensions in a unified framework.

CHAPTER 3

CONCLUSIONS

We examined the broad theme in asset pricing: how investors perceive different sorts of risks and how that would impact asset prices.

In Chapter One, we studied how social transmission, people's conversations along their social dimension, would influence their portfolio and trading choices, which would, in turn, influence asset prices. Individual investors, proxied by Reddit users, tend to follow biased presentations of investment results and, thus, over-buy high-crash-risk stocks.

In Chapter Two, we studied how cyberattacks on firms could update people's expectations about future probabilities for all the firms to get attacked in a certain year. That update of expectations has asset pricing consequences, as the most cyber-risky firms would get the most discount in their prices when the update happens.

These results also speak to the ongoing debate of whether strictly following the rational expectations paradigm is appropriate. One caveat of the studies here is that we do not have account-level trading data, so we do not perfectly observe how investors make portfolio choices in real-time. In future research, we hope to experiment with more settings and back out individual investors' realistic preference parameters and utility functions.

BIBLIOGRAPHY

Abreu, D. and Brunnermeier, M. K. (2003). Bubbles and crashes. *Econometrica*, 71(1):173–204.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56.

An, H. and Zhang, T. (2013). Stock price synchronicity, crash risk, and institutional investors. *Journal of Corporate Finance*, 21:1–15.

Andreou, P. C., Antoniou, C., Horton, J., and Louca, C. (2016). Corporate governance and firm-specific stock price crashes. *European Financial Management*, 22(5):916–956.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The Cross-Section of Volatility and Expected Returns. *Journal of Finance*, 61(1):259–299.

Atilgan, Y., Bali, T. G., Demirtas, K. O., and Gunaydin, A. D. (2020). Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. *Journal of Financial Economics*, 135(3):725–753.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of financial economics*, 99(2):427–446.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, 107(3):797–817.

Barber, B. M., Huang, X., Odean, T., and Schwarz, C. (2021). Attention induced trading and returns: Evidence from robinhood users. *Journal of Finance, forthcoming*.

Barber, B. M. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance*, 55(2):773–806.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2):785–818.

Barber, B. M., Odean, T., and Zhu, N. (2008). Do retail trades move markets? *The Review of Financial Studies*, 22(1):151–186.

Barber, B. M., Odean, T., and Zhu, N. (2009). Systematic noise. *Journal of Financial Markets*, 12(4):547–569.

Barberis, N. and Huang, M. (2008). Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, 98(5):2066–2100.

Bates, D. S. (2000). Post-'87 crash fears in the s&p 500 futures option market. *Journal of econometrics*, 94(1-2):181–238.

Beason, T. and Schreindorfer, D. (2022). Dissecting the equity premium. *Journal of Political Economy*, 130(8):2203–2222.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives*, 12(3):151–170.

Binfarè, M. (2019). The Real Effects of Risk Management Vulnerabilities: Evidence from Data Breaches. *Available at SSRN 3411553*.

Black, F. (1986). Noise. *The journal of finance*, 41(3):528–543.

Bollen, N. P. and Whaley, R. E. (2004). Does net buying pressure affect the shape of implied volatility functions? *The Journal of Finance*, 59(2):711–753.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Brunnermeier, M. K., Gollier, C., and Parker, J. A. (2007). Optimal beliefs, asset prices, and the preference for skewed returns. *American Economic Review*, 97(2):159–165.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2020). The structure of economic news. Technical report, National Bureau of Economic Research.

Callen, J. L. and Fang, X. (2015). Short interest and stock price crash risk. *Journal of Banking & Finance*, 60:181–194.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, 63(6):2899–2939.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

Chang, I.-C., Liu, C.-C., and Chen, K. (2014). The push, pull and mooring effects in virtual migration for social networking sites. *Information Systems Journal*, 24(4):323–346.

Chang, X. S., Chen, Y., and Zolotoy, L. (2016). Stock liquidity and stock price crash risk. *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming*.

Chen, A. Y. and Zimmermann, T. (2021). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*.

Chen, J., Hong, H., and Stein, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of financial Economics*, 61(3):345–381.

Chen, L., Pelger, M., and Zhu, J. (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75(3):1371–1415.

Cong, L. W., Harvey, C. R., Rabetti, D., and Wu, Z.-Y. (2023). An anatomy of crypto-enabled cybercrimes. Technical report, National Bureau of Economic Research.

Conrad, J., Kapadia, N., and Xing, Y. (2014). Death and jackpot: Why do individual investors hold overpriced stocks? *Journal of Financial Economics*, 113(3):455–475.

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, 267(5199):843–848.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990a). Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990b). Positive feedback investment strategies and destabilizing rational speculation. *the Journal of Finance*, 45(2):379–395.

Duffie, D. and Younger, J. (2019). Cyber runs. *Hutchins Center Working Paper*.

Eisenbach, T. M., Kovner, A., and Lee, M. J. (2020). Cyber risk and the us financial system: A pre-mortem analysis. *FRB of New York Staff Report*, (909).

Eisfeldt, A. L. and Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *Journal of Finance*, 68(4):1365–1406.

Fama, E. F. and French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds.

Journal of.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51(1):55–84.

Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fama, E. F. and French, K. R. (2020). Comparing cross-section and time-series factor models. *The Review of Financial Studies*, 33(5):1891–1926.

Fama, E. F. and MacBeth, J. D. (1973a). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Fama, E. F. and MacBeth, J. D. (1973b). Risk, return, and equilibrium: empirical tests. *Journal of Political Economy*, 81(3):607–636.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Florackis, C., Louca, C., Michaely, R., and Weber, M. (2022). Cybersecurity risk. *Review of Financial Studies*, forthcoming.

Foucault, T., Sraer, D., and Thesmar, D. J. (2011). Individual investors and volatility. *The Journal of Finance*, 66(4):1369–1406.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.

Gormley, T. A. and Matsa, D. A. (2011). Growing out of trouble? corporate responses to liability risk. *The Review of Financial Studies*, 24(8):2781–2821.

Graham, J. R. and Kumar, A. (2006). Do dividend clienteles exist? evidence on dividend preferences of retail investors. *The Journal of Finance*, 61(3):1305–1336.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, 70(3):393–408.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Han, B., Hirshleifer, D., and Walden, J. (2022). Social transmission bias and investor behavior.

Journal of Financial and Quantitative Analysis, 57(1):390–412.

Han, B. and Kumar, A. (2013). Speculative retail trading and asset prices. *Journal of Financial and Quantitative Analysis*, 48(2):377–404.

Harvey, C. R. and Siddique, A. (2000). Conditional skewness in asset pricing tests. *Journal of Finance*, 55(3):1263–1295.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of statistical learning*. Springer Science+Business Media New York.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Hirshleifer, D., Hou, K., Teoh, S. H., and Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, 38:297–331.

Hoberg, G. and Philips, G. M. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hu, D., Jones, C. M., Zhang, V., and Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Available at SSRN 3807655*.

Hutton, A. P., Marcus, A. J., and Tehranian, H. (2009). Opaque financial reports, r2, and crash risk. *Journal of financial Economics*, 94(1):67–86.

Jamilov, R., Rey, H., and Tahoun, A. (2020). The Anatomy of Cyber Risk. *London Business School Working Paper*.

Jang, J. and Kang, J. (2019). Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. *Journal of Financial Economics*, 132(1):222–247.

Jiang, H., Khanna, N., Yang, Q., and Zhou, J. (2020). The cyber risk premium. *Available at SSRN: https://ssrn.com/abstract=3637142 or http://dx.doi.org/10.2139/ssrn.3637142*.

Jin, L. and Myers, S. C. (2006). R2 around the world: New theory and new tests. *Journal of financial Economics*, 79(2):257–292.

Kamiya, S., Kang, J.-K., Kim, J., Milidonis, A., and Stulz, R. M. (2021). Risk management, firm reputation, and the impact of successful cyberattacks on target firms. *Journal of Financial Economics*, 139(3):719–749.

Kashyap, A. K. and Wetherilt, A. (2019). Some Principles for Regulating Cyber Risk. *AEA Papers and Proceedings*, 109:482–487.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. Technical report, National Bureau of Economic Research.

Kelley, E. K. and Tetlock, P. C. (2017). Retail short selling and stock prices. *The Review of Financial Studies*, 30(3):801–834.

Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Kim, J.-B., Li, L., Lu, L. Y., and Yu, Y. (2016). Financial statement comparability and expected crash risk. *Journal of Accounting and Economics*, 61(2-3):294–312.

Kim, J.-B., Li, Y., and Zhang, L. (2011). Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics*, 100(3):639–662.

Kim, J.-B. and Zhang, L. (2014). Financial reporting opacity and expected crash risk: Evidence from implied volatility smirks. *Contemporary Accounting Research*, 31(3):851–875.

Kim, Y., Li, H., and Li, S. (2014). Corporate social responsibility and stock price crash risk. *Journal of Banking & Finance*, 43:1–13.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Kopp, E., Kaffenberger, L., and Jenkinson, N. (2017). *Cyber risk, market failures, and financial stability*. International Monetary Fund.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2-3):221–247.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Li, X. and Wu, L. (2018). Herding and social media word-of-mouth: Evidence from groupon. *Forthcoming at MISQ*.

Lin, Z., Sapp, T. R., Ulmer, J. R., and Parsa, R. (2020). Insider trading ahead of cyber breach announcements. *Journal of Financial Markets*, 50:100527.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

McCrank, J. (2021). Factbox: The u.s. retail trading frenzy in numbers. *Thomson Reuters,*. URL: https://www.reuters.com/article/us-retail-trading-numbers-idUSKBN29Y2PW.

Michel, A., Oded, J., and Shaked, I. (2020). Do security breaches matter? The shareholder puzzle. *European Financial Management*, 26(2):288–315.

NBER (2021). Us business cycle expansions and contractions.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of financial economics*, 63(1):3–50.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition, 2011. *Linguistic Data Consortium, Philadelphia, PA, USA*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, 22(1):435–480.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of finance*, 52(1):35–55.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Van Buskirk, A. (2011). Volatility skew, earnings announcements, and the predictability of crashes.

Earnings Announcements, and the Predictability of Crashes (April 28, 2011).

Warren, P., Kaivanto, K., Prince, D., et al. (2018). Could a cyber attack cause a systemic impact in the financial sector? *Bank of England Quarterly Bulletin*, 58(4):21–30.

Welch, I. (2020). Retail raw: Wisdom of the robinhood crowd and the covid crisis. Technical report, National Bureau of Economic Research.

Xing, Y., Zhang, X., and Zhao, R. (2010). What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis*, pages 641–662.

Yan, S. (2011). Jump risk, stock returns, and slope of implied volatility smile. *Journal of Financial Economics*, 99(1):216–233.

Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278.

Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12.