Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus

Master's Thesis

Cyber risk and the cross-section of stock returns

Author
Daniel Celeny

Academic Supervisor
Prof. Pierre Collin-Dufresne

Company Supervisors
Dr. Alain Mermoud
Dr. Loïc Maréchal

A thesis submitted in fulfillment of the requirements for the Master degree in

Financial Engineering

August 25, 2023

Contents

List of Figures

List of Tables

Abstract

In this thesis, I extract firms' cyber risk with a machine learning algorithm measuring the proximity between their disclosures and a dedicated cyber corpus. This approach outperforms dictionary methods, is able to make use of the full disclosure and not only dedicated sections, and generates a cyber risk measure that is uncorrelated with other firms' characteristics. I find that a portfolio of US-listed stocks in the high cyber risk quantile generates an excess return of 18.72% p.a. Moreover, a long-short cyber risk portfolio has a significant and positive risk premium of 6.93% p.a., robust to all factors' benchmarks. Finally, using a Bayesian asset pricing method, I show that my cyber risk factor is the essential feature that allows any multi-factor model to price the cross-section of stock returns.

I appreciate helpful comments from seminar participants at the Cyber Alp Retreat 2023.

Chapter 1

Introduction

This thesis arises from a collaboration between EPFL and the Cyber-Defence Campus (CYD Campus). Founded in January 2009, the CYD Campus depends on armasuisse Science and Technology, the scientific branch of the federal office for defense procurement. More specifically, this study was carried out within the Technology Monitoring (TM) team of the CYD Campus, whose primary purpose is to provide an anticipation platform for cybersecurity technologies. TM employs both qualitative and quantitative approaches to identify emerging cyber technologies and firms. It does the former through scouting and the latter through the analysis of publicly available data. This research uses quantitative finance methods to contribute to the objectives of the CYD Campus.

One of the objectives of the CYD Campus is to guide stakeholders to make more informed decisions for investments and strategic procurement. Although the decision process has many ramifications, being able to get insights on the public equity markets is helpful to estimate the cost of firms associated with different types of risks, and hence the value of insurance that these firms should pay to mitigate these risks. In particular, I am interested in cyber risk.

The continuous digitization of our surroundings, coupled with the widespread utilization of Internet-of-Things devices and the intersection of geopolitical interests, fuels an ongoing surge in cyberattacks, accompanied by escalating costs. As Chuck Robbins, Chair and CEO at Cisco, put it,

> If it were measured as a country, then cybercrime — which was predicted to inflict damages totaling $6 trillion USD globally in 2021 — would be the world's third-largest economy after the U.S. and China.

As cyberattacks become more widespread and costly, cyber insurance contracts become vital both for public companies and governments, who

must assess the global cyber risk of the economy. These insurance contracts, however, need a thorough understanding of the systematic risks in the economy and the firm-level cyber risk. This is why Mario Greco, CEO of Zurich Insurance group, said in a recent interview that cyber-attacks are set to become "uninsurable" and called on governments to "set up private-public schemes to handle systemic cyber risks that can't be quantified, similar to those that exist in some jurisdictions for earthquakes or terror attacks"[1].

In this paper, I develop a method to quantify the cyber risk of a company based on its disclosures and investigate whether this risk is costly to firms in the form of a market risk premium on their stock returns. To do this, I collect financial fillings, monthly returns, and other firm characteristics for over 7,000 firms, listed on stock markets in the United States, between January 2007 and December 2022. I use a machine learning algorithm called "Paragraph Vector" in combination with the MITRE ATT&CK cybersecurity knowledgebase to score each firm's filing based on its cybersecurity content.

I find evidence that the cyber risk does not correlate with firm size, book-to-market ratio, beta, and other standard firms' characteristics known to help price stock returns. At the aggregated level, my measure shows a monotonic increasing trend, with a score moving from 0.51 to 0.54 out of one, whereas the cross-sectional distribution of that score is extremely narrow (standard deviation of 0.03). I compare my cyber risk measure across Fama-French 12 industries and find results supporting my intuition, with "Business Equipment" and "Telephone and Television Transmission" being the riskiest and "Oil and Gas" and "Utilities", the safest.

I find that the cyber risk sorted long-short portfolio, which invests in high cyber risk stocks and shorts low cyber risk stocks, has an average annual excess return of 6.93% and is statistically significant at the 10 or 5% level even when controlling for common risk factors. This portfolio performs especially well before the first release of a cyber risk factor on SSRN, with an average annual excess return of 11.88%, and is statistically significant at the 1% level.

I use asset pricing tests and find that the cyber risk generates a significant premium after controlling for market beta, book-to-market, size, momentum, operating profitability, and investment aggressiveness (see Fama and French, 2015). This performance shows up both in cross-section, with Fama and MacBeth (1973) regressions, and time series, with no significant joint alphas in Gibbons, Ross, and Shanken (1989) tests. Using the Bayesian approach of Barillas and Shanken (2018), I additionally show that the optimal subset of factors pricing stock returns always includes my cyber risk factor.

[1]Available at: https://www.ft.com/content/63ea94fa-c6fc-449f-b2b8-ea29cc83637d

Finally, I conduct tests to verify the robustness of my factor. First, I control that my baseline measure, revised at each new filing, captures the latent cyber risk and not the immediate effect of a cyberattack. To do so, I build a long-run cyber risk measure capturing the cumulative cyber risk effect. The results are virtually unchanged. Second, I control for the possibility that firms are providing cybersecurity and for which cyber risk occurrences might be positive. I also do not find any differences after that control.

The remainder of this work proceeds as follows. Chapter 2 introduces the previous literature and develops related hypotheses. Chapter 3 presents the data, and Chapter 4 the Methods. Chapter 5 details the results and Chapter 6 concludes.

Chapter 2

Literature review

2.1 Asset pricing

2.1.1 Factor models

Sharpe (1964), Lintner (1965) and Mossin (1966) independently introduced the Capital Asset Pricing Model (CAPM) for pricing individual assets or portfolios. According to this model, the expected return of an asset is determined by the risk-free rate of return, the asset's beta, and the market risk premium. Beta measures the sensitivity of the asset's returns to changes in the overall market. The CAPM has its limitations, however, and several other risk factors have been proposed to expand the CAPM model into a factor model.

More generally, factor models are used to model the covariance matrix of stock returns, by encoding information about a very large number of assets with a much smaller number of factors, such that the unexplained variations are uncorrelated. This relation can be written in a matrix form:

$$R^e = \alpha + \beta \boldsymbol{F} + \epsilon, \tag{2.1}$$

where R^e are the asset excess returns and $\boldsymbol{F}$ are the factor returns. Several factors were proposed, resulting in a "factor zoo". Most importantly, Fama and French (1992) show that two firm characteristics, other than market beta, predict returns: market capitalization and book-to-market ratio. The combination of these two factors with the market factor results in a 3-factor model. Carhart (1997) adds a momentum factor, following the work of Jegadeesh and Titman (1993), which is a portfolio that buys stocks that have performed well in the past year and sell stocks that have performed badly. Fama and French (2015) present an extension to their previous model with two new factors: investment and operating profitability, resulting in a five-factor model.

Following the publication of a large number of factors, Harvey, Liu, and Zhu (2016) study 316 factors. The authors claim that most research findings in financial economics are likely false, and many of the factors that their method deems statistically true have small Sharpe Ratios.

2.1.2 Asset pricing tests

Several asset pricing tests can be used to study factor models. Fama and MacBeth (1973) present a two-step regression method used to compute both the betas of the factors and their risk-premia. Gibbons et al. (1989) present a statistic used to test for portfolio efficiency. Their statistic can be generalized to test whether the pricing errors are jointly equal to zero in a model containing several traded factors. Another methodology, presented by Barillas and Shanken (2018), is used to compute the probability that a given factor model is best for pricing returns amongst all possible models spanned by the factors under consideration.

2.2 Vector representation of paragraphs

Le and Mikolov (2014) present an unsupervised algorithm called "Paragraph Vector" that can learn fixed-length vector representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Each piece of text represented by a dense vector can be used for text classification and sentiment analysis, for example. The advantage of this algorithm over other methods, such as bag-of-words, is that it learns the semantics of words and sentences. Lau and Baldwin (2016) perform a rigorous empirical evaluation of this algorithm and provide recommendations on hyper-parameter settings for general-purpose applications. Adosoglou, Lombardo, and Pardalos (2021) use the "Paragraph Vector" algorithm with financial filings (10-K statements). They construct portfolios based on the semantic differences between two consecutive financial reports of each firm. They find that cosine similarity is the most effective similarity measure to use with neural network embeddings, such as the ones obtained using the "Paragraph Vector" algorithm.

2.3 Cybersecurity investments

2.3.1 Optimal cybersecurity investments

Gordon and Loeb (2002) (hereafter, GL) study the optimal amount to invest to protect a given set of information. They use an economic model

to show that managers should focus on securing moderately vulnerable information sets and invest at most 37% of the expected loss due to cyber breaches. An extension to this model proposed by Gordon, Loeb, Lucyshyn, and Zhou (2015a) studies how externalities change the optimal investment and conclude that unless private sector firms consider the costs of breaches associated with externalities, they will underinvest in cybersecurity activities. In a further improvement to the GL model, Farrow and Szanton (2016) introduces a model where investments may reduce not only the probability of an attack but also the loss from an attack.

Using a real options method, Gordon, Loeb, Lucyshyn, and Zhou (2015b) explore the impact of information sharing on cybersecurity investments by showing that reduced uncertainty, resulting from information sharing, diminishes the value of the option to defer the investment. On the other hand, Willemson (2006) and Hausken (2006) disprove the 37% conjecture of the GL model, and present cases with required investment levels up to 100% of the expected loss.

2.3.2 Economic incentives for cybersecurity investments

Gordon, Loeb, Lucyshyn, and Zhou (2015c) assess the impact of government incentives and regulations designed to offset the tendency to underinvest in cybersecurity-related activities by firms. Their results show that the effectiveness of such incentives depends on the firm's usage of an optimal mix of cybersecurity inputs and their willingness to increase their investments in cybersecurity. Lelarge (2012) studies the outcomes of incentivizing agents of a large network towards better cybersecurity. He finds that for a large class of risks, only a small fraction of the expected loss should be invested. Security investments are always socially inefficient when agents are strategic, due to network externalities. He further shows that the alignment of incentives leads to a coordination problem and an equilibrium with a very high price of anarchy. Wang (2019) presents an analytical framework for firms to optimize their cybersecurity investment and cyber insurance program. He shows how private-sector efforts toward countering cybercrimes can reduce aggregate cyber loss and create economic value at the macro level. The private sector can reduce the total cybersecurity costs by pooling resources to pursue cyber offenders and seek loss recoveries actively. Small and medium-sized firms benefit most from additional security spending at a micro-level.

2.4 Cybersecurity costs

2.4.1 Direct estimations

Anderson et al. (2013) perform a systematic study of the costs of cybercrime. They disentangle direct, indirect, and defence costs and different types of cybercrimes. They find that traditional crimes that are conducted online, such as tax and welfare fraud, cost the typical citizen in the low hundreds of dollars per year. Transitional crimes, such as credit card fraud, cost a few dollars a year, while new crimes, such as the provision of botnets, cost tens of cents a year. Indirect and defence costs, however, are much higher for transitional and new crimes. They conclude that we should spend less in anticipation of cybercrime and more in response. Anderson et al. (2019) revisit the previous study. They observe that even though payment frauds have doubled over the seven years separating the initial studies, their average costs for the citizen have fallen. Their conclusion stays the same, that economic optimality would be in spending less on cyberattack prevention and more on response and law enforcement. By employing a Value at Risk (VaR) framework to analyze various cyber incidents and patterns in the financial sector worldwide, Bouveret (2018) documents significant cyber risk, revealing an average country-level loss of USD 97 billion and a VaR range of USD 47 to USD 201 billion, leading to the conclusion that potential aggregated losses in the financial sector far surpass the coverage capacity of the cyber insurance market by several orders of magnitude. Romanosky (2016) studies the composition and costs of cyber events. After analyzing a sample of over 12,000 cyber events, he finds that the cost distribution is heavily skewed, with an average cost of $ 6 million and a median cost of $ 170k (comparable to the firm's annual IT security budget). He concludes that with these relatively low costs, it may be that firms are engaging in a privately optimal level of security, and subsequently, firms are investing in only a modest amount of data protection.

2.4.2 Indirect estimations with municipal bonds

Andreadis, Kalotychou, Louca, Lundblad, and Makridis (2023) study the impact of information dissemination about cyberattacks through major news sources on municipalities' access to finance, focusing on the municipal bond market. They employ a differences-in-differences framework and find that both the cumulative number of cyberattacks covered by county-level news articles and the corresponding number of county-level cyberattack news articles have a significant adverse effect on municipal bond yields. A 1% increase in the number of cyberattacks covered by news articles leads to an increase in offering yields ranging from

3.7 to 5.9 basis points, depending on the level of attack exposure (number of major cyberattack news in the county). Jensen and Paine (2023) perform a similar analysis, using data about municipal IT investment, ransomware attacks and bonds. They find no immediate effect on bond yields of hacked towns in a 30-day window around a hack. In the 24 months following a ransomware attack, they find that the municipal bond yields gradually decline and IT spending increases. They argue that the declining bond yields are driven by a decrease in the cyber risk of the town as a result of the increase in IT spending.

2.4.3 Indirect estimations with stock price reactions

Gordon, Loeb, and Zhou (2011) study the impact of information security breaches on stock returns by computing the cumulative abnormal returns on a three-day event window centered on newspaper reports of cybersecurity incidents. They find that news about information security breaches had a statistically significant effect on the stock returns of publicly traded firms. They also show that there has been a significant downward shift in the impact of security breaches in the post-9/11 period, as they have become less costly and investors view them as a corporate "nuisance" rather than a potentially serious economic threat. In a similar study, Campbell, Gordon, Loeb, and Zhou (2003) find a highly significant negative market reaction for information security breaches involving unauthorized access to confidential data but no significant reaction when the breach does not involve confidential information. Johnson, Kang, and Lawson (2017) also study cumulative abnormal returns around cybersecurity events. They show that publicly traded firms in the U.S. lost, on average, 0.37% of their equity value when a data breach occurs. The biggest decline of equity value (3% on average) is due to payment card fraud, when the card breaches are larger than the average.

Lending, Minnick, and Schorno (2018) study the relationship between corporate governance and the probability of data breaches. They measure the changes in stock returns following data breaches and find that the financial impact of a breach is visible in the long term, as data-breach firms have –3.5% one-year buy-and-hold abnormal returns. They also find that banks with breaches have significant declines in deposits and non-banks have significant declines in sales in the long run. Tosun (2021) studies how financial markets react to unexpected corporate security breaches in the short and long run. He finds that the market reaction in terms of trading volume, liquidity, and sell pressure anticipates negative changes in stock prices, which turn out to be significant and negative only the day after security breaches are publicly announced. He also finds that cyberattacks affect firms' policies in the long run. He concludes that security breaches represent unexpected negative shocks

to firms' reputations. Kamiya, Jun-Koo, Jungmin, Milidonis, and Stulz (2021) also find evidence of a reputation loss for target firms, in the form of a decrease in credit ratings or decreased sales growth.

2.4.4 Indirect estimations with disclosures

Gordon, Loeb, and Sohail (2010) assess the market value of voluntary information security disclosures of firms, using a sample of 1,641 disclosing and 19,266 non-disclosing firm-year observations. They argue that voluntary disclosures pertaining to information security could serve to mitigate potential litigation costs and lower the firm's cost of capital by reducing the information asymmetry between a firm's management and its investor. They find a positive association of the voluntary disclosure variable with firm value, and the bid-ask spread for firms that provide voluntary disclosures of information security is statistically lower than for firms not providing such disclosure. Hilary, Segal, and Zhang (2016) also study cyber risk disclosures. They find that the market reaction to cyber breaches is statistically significant but economically limited.

Florackis, Louca, Michaely, and Weber (2023) build a text-based cyber risk measure using a section of 10-K statements called "Item 1.A Risk Factors". They extract cyber risk-related sentences from this section of the statements using a list of keywords and restrict the analysis to these sentences. They consider recently hacked firms as a training sample and compute the cybersecurity exposure of firms as the average similarity between the bag-of-words representation (vector of the number of occurrences of each word in their dictionary) of the firm's cybersecurity sentences and the cybersecurity sentences of the training sample. They find that stocks with high exposure have higher returns on average but perform worse in periods of cyber risk. Their value-weighted long-short portfolio has an average monthly excess return of 0.6% (based on tercile portfolios). Jamilov, Rey, and Tahoun (2021) perform a similar analysis using quarterly earnings calls. They construct the cyber risk measure using the frequency of cybersecurity-related keywords in the earnings calls. They build a cybersecurity factor by first computing the monthly average cyber risk score of the subset of firms with non-zero scores and then fitting an AR(1) model to the time series and extracting the residuals. This factor captures the shocks to cyber risk. They find a factor structure in the firm-level measure of cybersecurity, that is the long-short portfolio built on cybersecurity beta sorted portfolios has an average annual return of -3.3% (the sign is due to the fact that the factor captures shocks to cybersecurity).

I am only aware of the two studies above that focus on cyber risk and its factor structure using disclosures. However, both studies use a dic-

tionary approach that leaves many firm-year observations with a cyber risk of zero (71% of firms in 2007 in the case of Florackis et al. (2023) and over 98% of earnings calls in 2007 in the case Jamilov et al. (2021)). Furthermore, this approach does not take into account the context of the keywords only their presence in the disclosures. To fill this gap, I use the "Paragraph Vector" algorithm to build a cyber risk measure using firms' 10-K statements. Hence, I define my null hypotheses as follows:

- H_a The cyber risk is not priced in the cross-section of stock returns
- H_b The cyber risk factor is subsumed by other factors

Chapter 3

Data

3.1 Market data

I download public equity data from Wharton Research Data Services (WRDS)[2], in which I use the data from the Center for Research in Security Prices (CRSP)[3] and S&P Global Market Intelligence's Compustat database[4]. I report the list of variables in Table A.3. I develop a Python script that queries all available information from WRDS' API and filters the firms based on the existence of a 10-K filing with the SEC (see Chapter 3.2 below) so that all of the retained firms have at least one 10-K statement available. I extract monthly stock returns and financial ratios for 7,059 firms, between January 2007 and December 2022. I depict the industry distribution of these firms, using the Fama-French 12 industry classification in Figure 3.1.

I also download the one-month Treasury bill rate and returns on the market, book-to-market (HML), size (SMB), momentum (MOM), investment (CMA) and operating profitability (RMW) factors from the Kenneth French data repository[5].

3.2 10-K statements

10-K statements are financial filings submitted by publicly traded companies to the U.S. Securities and Exchange Commission (SEC) annually. They contain information such as companies' financial statements, risk factors, and executive compensation. I use 10-K statements to build a cyber risk measure (detailed in Chapter 4).

[2]Available at: https://wrds-www.wharton.upenn.edu/
[3]Available at: https://crsp.org/
[4]Available at: https://www.marketplace.spglobal.com/en/datasets/compustat
[5]Available at: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

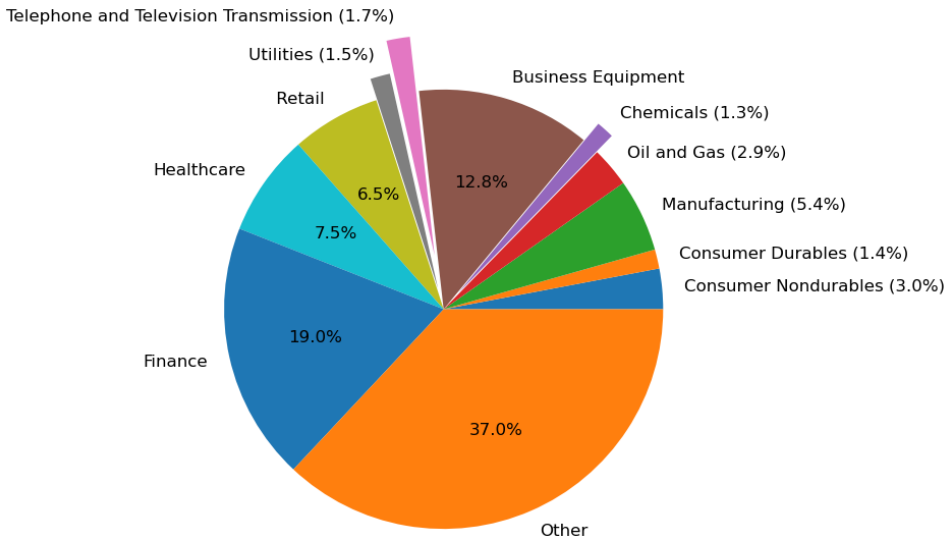

Figure 3.1: **Industry distribution**

Distribution of firms in the 12 Fama-French industries. Standard Industrial Classification (SIC) codes are obtained from CRSP. The conversion table, from SIC to 12 Fama-French industries, is available on the Kenneth French data repository.

To download these statements, I use the index files from the SEC's Edgar archives[6]. These index files contain information about all the documents filed by all firms for a specific quarter. Each line of the index file corresponds to a document and is structured as follows:

CIK | Company Name | Form Type | Date Filed | Filename

where Filename is the URL under which an HTML version of the document is available. To identify firms, I use their Central Index Key (CIK), which is a number used by the SEC to identify corporations and individuals who have filed disclosures. I develop a Python script that goes through these index files and identifies URLs that correspond to 10-K statements, using the Form Type entry. These URLs are then matched to one of the 7,059 firms mentioned in Chapter 3.1, using the CIK entry. 60,470 10-K statements are identified, which corresponds to 8.6 statements per firm on average. Figure 3.2 shows the number of 10-Ks filed per year. This number increases from 3,301 in 2007 to 5,370 in 2022.

[6]Available at: https://www.sec.gov/Archives/edgar/full-index/

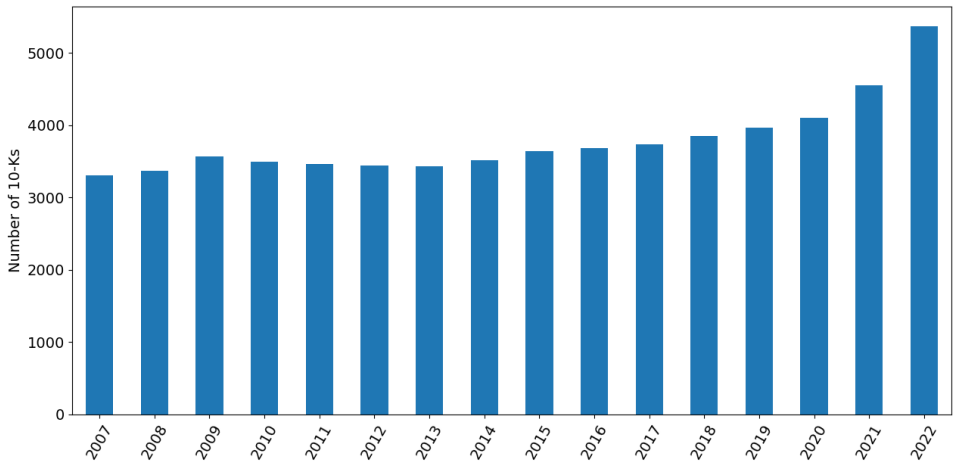

Figure 3.2: **Number of 10-Ks per year**

Number of companies, in the study sample, that have filed a 10-K statement in a given calendar year.

3.3 Cybersecurity tactics

I use the MITRE ATT&CK[7] cybersecurity knowledge base as a reference for cybersecurity descriptions. This knowledge base was created in 2013 to document cybersecurity tactics, techniques, and procedures used by adversaries against particular platforms, such as Windows or Google Workspace. Figure 3.3 illustrates the structure of the knowledge base. Each sub-technique has a short description describing it. Table 3.1 shows two sub-technique descriptions from the knowledge base.

There are a total of 14 tactics: reconnaissance, resource development, initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, command and control, exfiltration and impact. There are 785 sub-techniques across all tactics, all of which are used in Chapter 4.

[7]Available at: https://attack.mitre.org/

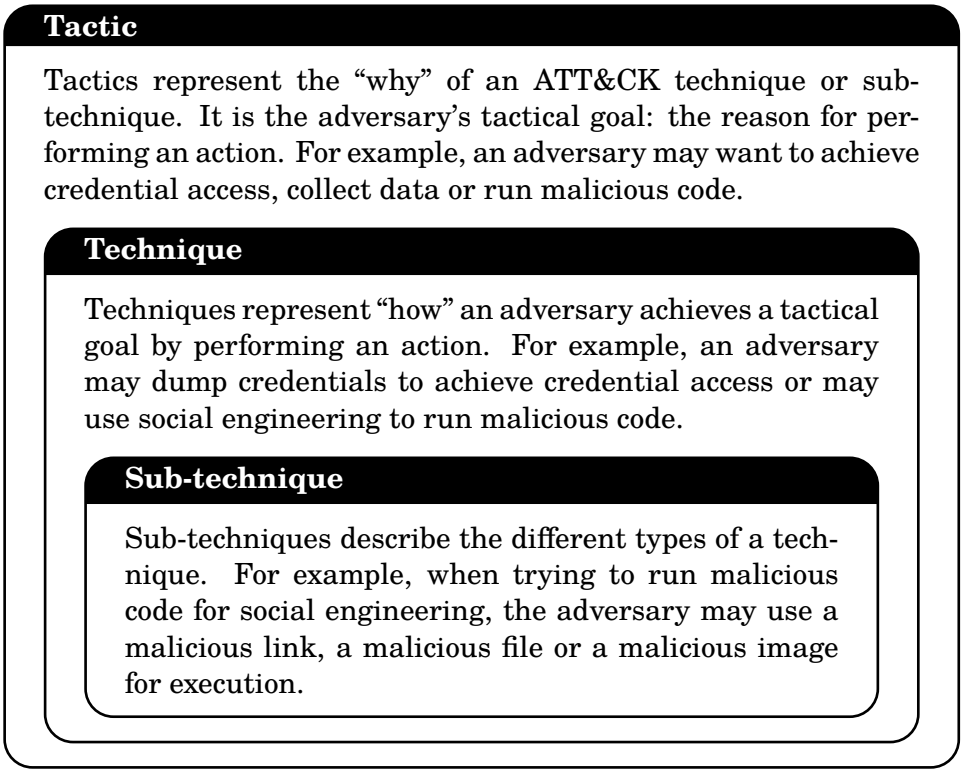

Figure 3.3: **Structure of the tactic descriptions on MITRE ATT&CK**

		Description
Tactic	Credential Access	Adversaries may forge web cookies that can be used to gain access to web applications or Internet services. Web applications and services (hosted in cloud SaaS environments or on-premise servers) often use session cookies to authenticate and authorize user access.
Technique	Forge Web Credentials	
Sub-technique	Web Cookies	
Tactic	Reconnaissance	Adversaries may gather employee names that can be used during targeting. Employee names can be used to derive email addresses as well as to help guide other reconnaissance efforts and/or craft more-believable lures.
Technique	Gather Victim Identity Information	
Sub-technique	Employee Names	

Table 3.1: **Examples of sub-technique descriptions from MITRE ATT&CK**

Chapter 4

Methodology

4.1 Models and assumptions

4.1.1 Text preprocessing

10-K statements can be downloaded from the SEC Archives as HTML files (as explained in Chapter 3). I use the BeautifulSoup[8] Python library to extract the usable text from these files. I remove the punctuation and numbers and I set all letters to lowercase. Given the resulting texts, I develop a Python script that uses the wordfreq[9] and NLTK[10] libraries to divide the text into sentences, remove stop-words (“the”, “is”, “and”,...) and remove the most common words of the English language. Since these words appear frequently in the texts, removing them allows us to focus on important words, that are related to cybersecurity for example.

After pre-processing, the average length of the cybersecurity subtechnique descriptions from MITRE ATT&CK is close to 40 words (≈ 39.7). Based on this number, I develop a Python algorithm to merge consecutive sentences from 10-K statements, into paragraphs that have an average length of close to 40 words, after pre-processing. On average, I obtain 44 words per paragraph and 638 paragraphs per 10-K statement, with standard deviations of 2.6 words per paragraph and 304 paragraphs per 10-K statement.

4.1.2 Paragraph Vector

The Paragraph Vector model, proposed by Le and Mikolov (2014), is an extension of the word2vec model (Mikolov, Chen, Corrado, and Dean (2013)). The Paragraph Vector model aims to learn fixed-length vector

[8]Available at: https://www.crummy.com/software/BeautifulSoup/
[9]Available at: https://pypi.org/project/wordfreq/
[10]Available at: https://www.nltk.org/

representations from variable-length pieces of text. The main advantage of this model over other methods, such as bag-of-words, is that semantically similar paragraphs are mapped close to each other in the vector space.

There are two versions of the model: a distributed memory model (DM) and a distributed bag-of-words model (DBOW). In the distributed memory model, the algorithm trains to get both word vectors and paragraph vectors. During training, the concatenation or the average of the paragraph vector and the vector representation of context words are used to predict another word in the paragraph. In the distributed bag-of-words model, the paragraph vector is trained to predict words in a window sampled from the paragraph. Word vectors are not trained in this version. Figure 4.1 illustrates the two models.

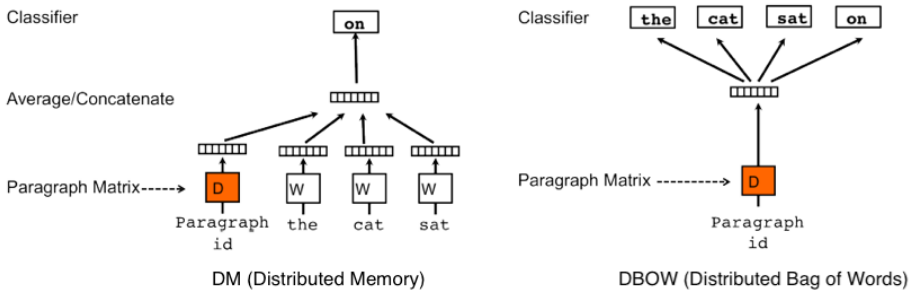

Figure 4.1: **"Paragraph Vector" model versions**

The images were taken from Le and Mikolov (2014).

Both models are unsupervised, as the paragraph vectors are learned from unlabelled data.

I use the implementation by Gensim called doc2vec[11]. To train the model, I use the paragraphs from 10-K statements filed in 2007 as well as the 785 sub-technique descriptions from MITRE, which together amount to more than 1.7 million training paragraphs. Using this training sample, I train DM and DBOW doc2vec models with various vector dimensions, epochs, and window sizes. The baseline for the hyperparameters is taken from Lau and Baldwin (2016) (see table A.1).

To choose the best model, I compute the vector representations of the paragraphs from 10 randomly chosen 10-K statements from 2008 (validation sample) using each model and compare the highest-scoring paragraphs between the models (the scoring algorithm is explained below). I choose the best model as the one where the proportion of the highest-scoring paragraphs that are cybersecurity-related is the highest. Table 4.1 presents the parameters of the best-performing doc2vec

[11]Available at: https://radimrehurek.com/gensim/models/doc2vec.html

model, which is the one used for the remainder of this study. Table A.2 presents the top-scoring paragraphs from the validation sample (after pre-processing).

Method	Training Size	Vector Size	Window Size	Min Count	Sub-Sampling	Negative Sampling	Epoch
DBOW	1.7M	200	15	5	10^{-5}	5	50

Table 4.1: **doc2vec parameters**

Parameters of the chosen doc2vec model. DBOW stands for distributed bag-of-words.

4.1.3 Cosine similarity

I use cosine similarity to measure the distance between the vector representations of two paragraphs, that is, the cosine of the angle between the two vectors. As explained in Adosoglou et al. (2021), cosine similarity is the most effective similarity measure as the orientation of the embedding vectors is more stable than their magnitude due to the random initialization of the weights of the neural networks.

Cosine similarity is a number between -1 and 1. The closer the vectors, the higher the value. As detailed in Chapter 4.1.4, I only use the positive cosine similarities and set the negative similarities to zero. The similarity between two paragraphs, with vector representations v_1 and v_2, is therefore computed as $sim = max(0, \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|})$.

4.1.4 Cyber risk score

The cyber risk score is based on the cosine similarities with the cybersecurity descriptions from MITRE ATT&CK. I first compute the vector representation of every paragraph of every 10-K statement using the trained doc2vec model. I also compute the vector representation of every sub-technique description from MITRE ATT&CK. Next, I compute the cosine similarity of each paragraph from the 10-K statements with each of the MITRE descriptions. This gives 785 similarities for each paragraph from the 10-K statements. The cyber risk score of a paragraph is the maximum value out of those 785 similarities. Finally, I compute the score of a 10-K statement as the average score of the 1% of its highest-scoring paragraphs.

This algorithm is based on the assumption that an average 10-K statement has at most six or seven cyber risk-related paragraphs, representing on average 1% of the paragraphs (there are 638 paragraphs per 10-K statement on average). This method has several advantages. First, taking a percentage of the total number of paragraphs, as opposed to a fixed number of paragraphs, makes it possible to have a meaningful comparison between 10-K statements that are much shorter or much

longer than the others. Second, considering only the highest-scoring paragraphs makes the cyber risk score of the 10-K statement only dependent on paragraphs that are most likely to be cyber risk-related.

As Chapter 4.1.3 mentions, the cosine similarity takes values between -1 and 1. I only consider the positive values for several reasons. First, a paragraph with a meaning "opposite" to cybersecurity is not intuitive. Upon inspecting the paragraphs with negative values in the validation sample, I cannot uncover meaningful differences between paragraphs with negative scores and those with scores close to zero. Furthermore, only considering positive similarities guarantees that the cyber risk scores are between 0 and 1, making them comparable to the ones obtained using dictionary methods such as in Jamilov et al. (2021).

Figure 4.2 shows the distribution of the cyber risk scores of the paragraphs from the 10-K statements of Meta Platforms, Inc. and Tesla, Inc. filed in 2022. The paragraphs in red are the top 1% of paragraphs with the highest cyber risk scores. I compute the 10-K cyber risk scores as the average score of this highest percentile. This yields a score of 0.605 for META and 0.563 for TSLA.

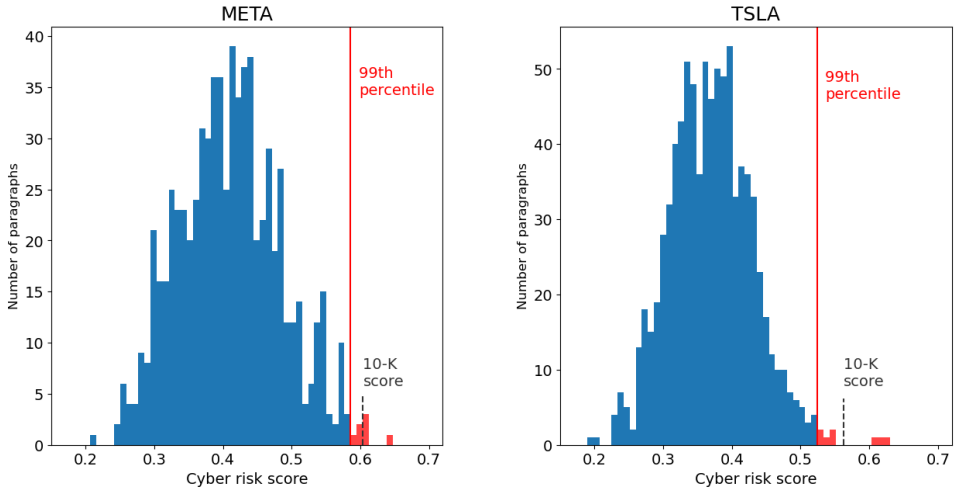

Figure 4.2: **Paragraph level score distributions for Meta Platforms and Tesla**

The paragraphs are the ones from the 10-Ks filed in 2022. The paragraphs within the top 1% of cyber risk scores are in red.

4.2 Asset pricing tests

4.2.1 Fama and MacBeth (1973)

I implement the methodology from Fama and MacBeth (1973) as follows. First, I estimate security betas using time series regressions with 3-year rolling windows. This corresponds to the following regression:

$$R_i = \alpha_{i,t} + \sum_k \beta_{i,t}^k F_k + \epsilon_{i,t}, \quad \forall i \tag{4.1}$$

where $\beta_{i,t}^k$ is the Ordinary Least Squares beta of asset i on factor k for the 3-year period ending on date t.

Next, I sort firms into 20 value-weighted portfolios based on their cyber risk beta. I compute the factor exposures of the portfolios as

$$\beta_p^k = \sum_{i=1}^{N} x_{i,p} \beta_i^k \tag{4.2}$$

where x_{ip} is the weight of asset i in portfolio p. As explained in the original paper, the betas of portfolios can be much more precise estimates of the true betas than the betas of individual securities if the errors in the betas of individual securities are substantially less than perfectly correlated. I standardize the portfolio betas for economic interpretation.

Finally, I estimate gammas using cross-sectional regressions of the portfolio returns on their lagged factor exposures. This corresponds to the following regression:

$$R_{p,t} = \gamma_t^0 + \sum_k \gamma_t^k \beta_{p,t-1}^k + \epsilon_{p,t}^*, \quad \forall t \tag{4.3}$$

where γ_t^k is the risk premium of factor k. I compute the average risk premiums over time.

4.2.2 Gibbons et al. (1989)

Gibbons et al. (1989) presents a statistic (GRS) to test for portfolio efficiency, based on the following regression:

$$R_{i,t} = \alpha_{i,p} + \beta_{i,p} R_{p,t} + \epsilon_{i,t}, \quad \forall i \tag{4.4}$$

The null hypothesis of the GRS test is $H_0 : \alpha_{i,p} = 0$. This statistic can be generalized to test whether the pricing errors are jointly equal to zero when using a model with several traded factors. As presented in Cochrane (2005), the regression equation is now

$$R_i = \alpha_i + \sum_k \beta_i^k F_k + \epsilon_i, \tag{4.5}$$

where R_i are the returns of portfolio i and F_k are the returns of factor k. Similarly to the Fama-Macbeth methodology, this methodology uses portfolios rather than individual securities. Hence, I build 20 portfolios on a factor beta, as explained in Chapter 5.4.

After computing the regressions from equation 4.5, I compute the GRS test statistic as,

$$\frac{T-N-K}{N}\frac{\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}}{1+\hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}} \sim F_{N,T-N-K}, \tag{4.6}$$

where T is the number of time periods, N is the number of portfolios, K is the number of factors, $\hat{\Sigma}$ is the residual covariance matrix, $\hat{\alpha}$ is the vector or alphas, $\hat{\mu}$ is the vector of average factor returns and $\hat{\Omega}$ is the covariance matrix of factors.

I implement this GRS test and compare two model specifications, the five-factor model from Fama and French (2015) and the same five factors plus the cyber risk factor.

4.2.3 Barillas and Shanken (2018)

I implement the Bayesian approach of Barillas and Shanken (2018). Using this method, it is possible to compute the probability that a given factor model is best for pricing factor returns. The method does not presume that any model under consideration exactly satisfies the requirement that all alphas are zero, as it is possible that some relevant factors have not been identified. The approach compares the relative success of the models in predicting the data.

The method is based on Barillas and Shanken (2017). The method looks at the extent to which a model prices the factors left out and not the extent to which the model prices test assets.

The unrestricted factor model is

$$R_t = \alpha + \beta F_t + \epsilon_t, \;\; \epsilon_t \sim N(0,\Sigma), \tag{4.7}$$

and the null hypothesis is $H_0 : \alpha = 0$. The prior for α is concentrated at zero under the null hypothesis. Under the alternative, they assume a multivariate normal informative prior for α: $P(\alpha|\beta,\Sigma) = MVN(0, k\Sigma)$, where k reflects the beliefs about the potential magnitude of deviations from the expected return relation. By assumption, all the models contain the market factor.

The marginal likelihood of a model is given by:

$$ML = ML_U(f|Mkt) \times ML_R(f^*|Mkt, f) \times ML_R(r|Mkt, f, f^*), \tag{4.8}$$

where ML_U is the unrestricted regression marginal likelihood, ML_R is the restricted regression marginal likelihood (α constrained to zero), f

are the included factors and f^* are the excluded factors. The $ML_U(X|Y)$ notation assumes the following regression equation:
$X_t = \alpha + \beta Y_t + \epsilon_t$. The unrestricted and restricted regression marginal likelihoods are given by,

$$\begin{aligned} ML_U &= |F'F|^{-N/2}|S|^{-(T-K)/2}Q \\ ML_R &= |F'F|^{-N/2}|S_R|^{-(T-K)/2}, \end{aligned} \tag{4.9}$$

where $|S|$ and $|S_R|$ are the determinants of the $N \times N$ cross-product matrices of the Ordinary Least Squares residuals, T is the number of periods, K the number of factors, and N the number of portfolios. The scalar Q is given by

$$Q = \left(1 + \frac{a}{a+k}(W/T)\right)^{-(T-K)/2} \left(1 + \frac{k}{a}\right)^{-N/2} \tag{4.10}$$

where $a = {}^{(1+Sh(F)^2)}/_{T}$, $k = {}^{(Sh^2_{max}-Sh(F)^2)}/_{N}$, W is the GRS F-statistic times ${}^{NT}/_{(T-N-K)}$ and $Sh(F)^2 = \mu'\Omega^{-1}\mu$ the squared sample Sharpe Ratio. Under the alternative prior, k is the expected increment to the squared Sharpe ratio from the addition of one more factor. Sh_{max} is the maximum expected Sharpe ratio. Barillas and Shanken (2018) take $Sh_{max} = 1.5 \times Sh_{Mkt}$, which corresponds to a square root of the prior expected squared Sharpe ratio for the all factors-tangency portfolio 50% higher than the market Sharpe ratio. They call this value the prior multiple. Similarly, I use 1.5 as the baseline value for the prior multiple but experiment with several values as in the original paper. The posterior probabilities, conditional on the data D, are given by Bayes' rule,

$$P(M_j|D) = \frac{ML_j \times P(M_j)}{\sum_i ML_i \times P(M_i)}, \tag{4.11}$$

where $P(M_j)$ is the prior probability of the model. Barillas and Shanken (2018) use uniform prior probabilities to avoid favoring one model over another. Hence, they cancel out in the division and can be omitted. $ML_R(r|Mkt, f, f^*)$, in equation 4.8, is the same for all combinations of f and f^*. Hence, it cancels out in the division and can also be omitted.

Following the methodology of Barillas and Shanken (2018), I compute the posterior probabilities for each month from January 2010 until December 2022. For each computation, I use all of the data available from January 2009 until the given time.

Chapter 5

Results

5.1 Cyber risk measure

Table 5.1 presents descriptive statistics of the cyber risk measure and various firm characteristics. The average cyber risk is 0.52, and its distribution is positively skewed, meaning there are more very high-risk firms than very low-risk ones. Overall, the cyber risk distribution is narrow with a standard deviation of 0.03 and a spread between the top and bottom percentiles of 0.14. The correlation coefficients between my cyber risk measure and firms' characteristics are small, except for Tobin's Q (0.23) and the Firm age (-0.17). The latter coefficient is consistent with the view that older firms are less subject to cyberattacks since their core businesses are less likely to be IT-related. Given that all other coefficients are below 0.15 in absolute value, I am confident that my measure is orthogonal to other characteristics known to price stock returns. I also compute the correlation coefficient between my measure and that of Florackis et al. (2023) and obtain 0.34.[12]

5.1.1 Time series and industry properties

Figure 5.1 presents the cross-sectional average cyber risk for every year in the study sample. I observe a monotonic positive time trend in line with the results of Florackis et al. (2023) and Jamilov et al. (2021).

Figure 5.2 shows the average cyber risk by industry, using the Fama-French 12 industry classification. Industries that rely on technology systems such as "Business Equipment" and "Telephone and Television Transmission" have high cyber risk while industries such as "Oil and Gas" and "Chemicals", that traditionally rely less on technology systems

[12]The Florackis et al. (2023) data is available at:
https://alucutac-my.sharepoint.com/personal/christodoulos_louca.

	Mean	SD	P1	P25	P50	P75	P99	Correlation with cyber risk
Cyber risk	0.52	0.03	0.47	0.50	0.52	0.54	0.61	-
Firm Size (ln)	20.18	2.39	13.15	18.53	20.25	21.86	25.46	-0.10
Firm Age (ln)	2.70	1.06	-0.88	2.21	2.93	3.41	4.07	-0.17
ROA	-0.11	0.47	-2.57	-0.07	0.02	0.07	0.36	-0.05
Book to market ratio	0.68	1.15	0.02	0.24	0.46	0.81	4.42	-0.12
Tobin's Q	2.20	2.14	0.58	1.09	1.50	2.37	12.15	0.23
Market Beta	1.20	0.84	-1.01	0.71	1.13	1.60	3.90	0.00
Intangibles/Assets	0.17	0.21	0.00	0.00	0.07	0.27	0.78	0.14
Debt/Assets	0.53	0.28	0.06	0.32	0.52	0.70	1.48	-0.09
ROE	-0.08	0.61	-2.96	-0.08	0.07	0.15	0.88	-0.06
Price/Earnings	1.55	112.17	-511.4	-4.44	12.57	23.82	294.46	-0.01
Profit Margin	-0.38	5.53	-25.20	0.21	0.36	0.57	0.94	0.00
Asset Turnover	0.92	0.74	0.01	0.39	0.76	1.26	3.54	-0.03
Cash Ratio	1.85	3.41	0.01	0.23	0.65	1.81	18.20	0.11
Sales/Invested Capital	1.54	1.59	0.01	0.56	1.08	1.94	8.88	-0.01
Capitalization Ratio	0.30	0.32	0.00	0.02	0.24	0.47	1.54	-0.10
R&D/Sales	0.67	4.21	0.00	0.00	0.00	0.08	19.40	0.03
ROCE	0.00	0.45	-1.97	-0.02	0.09	0.17	0.95	-0.07

Table 5.1: **Descriptive statistics of the cyber risk measure and firm characteristics**

Firm-level characteristics are winsorized at the 1st and 99th percentile (by year). The characteristics are defined in Table A.3.

have lower scores. Nonetheless, the variation of cyber risk across industries remains limited similar to the overall cyber risk distribution.

5.1.2 Determinants of firm-level cyber risk

To investigate the dependence of cyber risk on firm characteristics, I perform two regressions, presented in Table 5.2. In model 1, I control for year- and firm-fixed effects, and in model 2, I control for year- and industry-fixed effects. In both models, firm age has a statistically significant negative coefficient at the 1% level, implying that younger firms have higher cyber risk. The book-to-market coefficient is negative and significant at the 1% level, meaning that value firms have a lower cyber risk than growth firms. In model 2, the intangible assets to total assets coefficient is positive and statistically significant at the 1% level, which supports the view that firms with more intangible assets, such as patents or software, have a higher cyber risk. The R-squared is low for both models, showing that cyber risk can not be readily explained by firm characteristics.

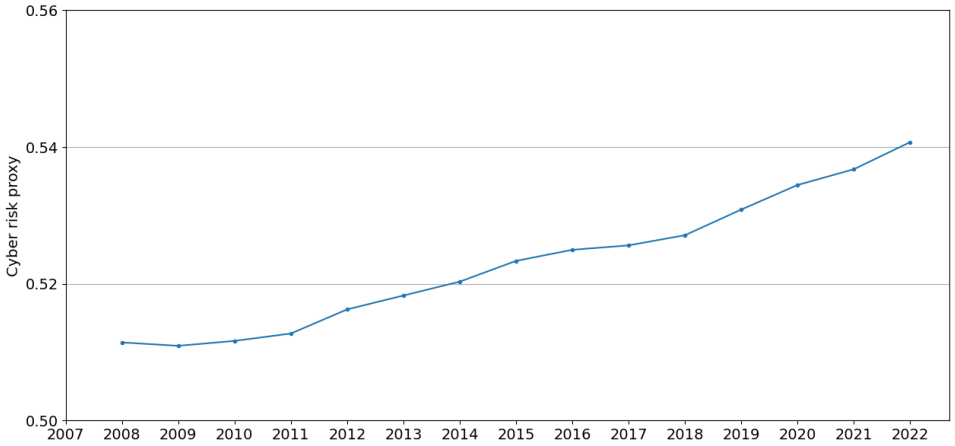

Figure 5.1: **Evolution of the average cyber risk across all firms**

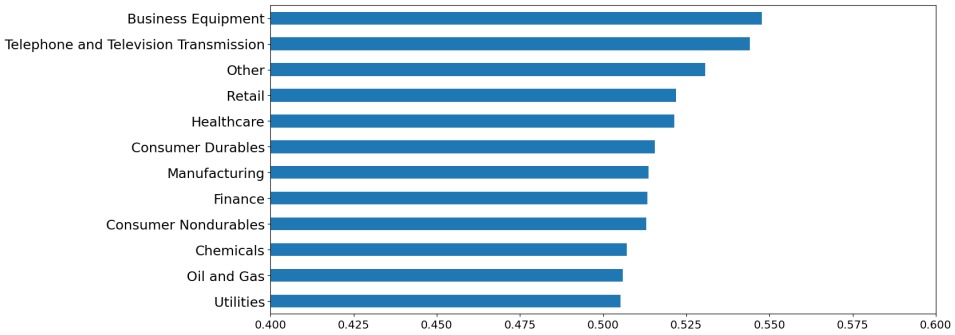

Figure 5.2: **Average value of the cyber risk across industries**

Firms are classified into industries using the Fama-French 12 industry classification. Standard Industrial Classification (SIC) codes are obtained from CRSP. The conversion table, from SIC codes to the 12 Fama-French industries, is available on the Kenneth French data repository.

Dependent variable: Firm-level indicator of cyber risk		
	Model 1	Model 2
Constant	**-0.416*****	**-0.738*****
	[-24.89]	[-14.22]
Firm Size (ln)	**0.019**	**0.024**
	[0.56]	[1.22]
Firm Age (ln)	**-0.114*****	**-0.211*****
	[-3.91]	[-12.09]
ROA	**0.057**	**0.0321****
	[0.79]	[2.27]
Book to Market	**-0.023*****	**-0.066*****
	[-4.82]	[-3.67]
Tobin's Q	**0.019*****	**0.112*****
	[2.84]	[7.87]
Market Beta	**-0.009**	**-0.013**
	[-1.54]	[-1.31]
Intangibles/Assets	**-0.026****	**0.082*****
	[-2.04]	[5.51]
Debt/Assets	**-0.032****	**0.032**
	[-2.49]	[1.21]
ROE	**0.002**	**-0.009**
	[0.37]	[-0.72]
Price/Earnings	**0.005**	**0.002**
	[1.20]	[0.29]
Profit Margin	**0.006**	**0.048*****
	[1.18]	[3.98]
Asset Turnover	**-0.014**	**-0.135*****
	[-0.67]	[-4.49]
Cash Ratio	**0.001**	**0.019**
	[0.09]	[1.19]
Sales/Invested Capital	**0.008**	**0.104*****
	[0.54]	[3.80]
Capital Ratio	**0.001**	**-0.191*****
	[0.02]	[-8.49]
R&D/Sales	**-0.001**	**-0.003**
	[-0.22]	[-0.30]
ROCE	**0.005**	**0.000**
	[0.71]	[0.01]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	27760	27760
R-squared	0.2944	0.3921

Table 5.2: **Determinants of firm-level cyber risk**

Results of regressions of the cyber risk on firm characteristics. t-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The characteristics are defined in Table A.3.

5.2 Univariate portfolio sorts

I sort firms into portfolios based on their cyber risk and study the returns of the portfolios. More precisely, I assign firms to five portfolios based on the cyber risk score of their most recent 10-K statement. I rebalance the portfolios quarterly to allow for listings and delistings and incorporate information from new 10-K statements. I build five value-weighted portfolios, where Portfolio 1 (5) is the low (high) cyber risk portfolio.

5.2.1 Full sample

I track the performance of the portfolios from January 2009 until December 2022. Figure 5.3 shows the evolution of the cumulative returns of the five portfolios and the market portfolio. I observe that the higher the cyber risk of the portfolio, the higher the cumulative returns. Portfolio 5 significantly outperforms the market.

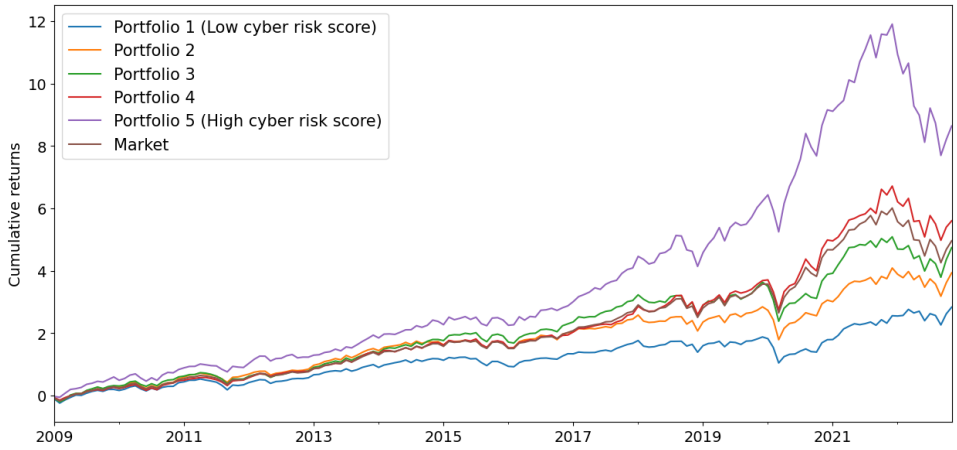

Figure 5.3: **Cyber risk sorted portfolio cumulative returns**

Firms are sorted into value-weighted portfolios based on their cyber risk. The portfolios are rebalanced quarterly. "Market" refers to the market portfolio obtained from the Kenneth French data repository.

Table 5.3 presents the excess returns and alphas of the portfolios with respect to three traditional factor models. The average monthly portfolio excess returns increase monotonically from 0.88% to 1.44%, from the low to high cyber risk portfolios. The long-short portfolio, going long in Portfolio 5 and short in Portfolio 1, has statistically significant excess returns and alphas, even when controlling for the Fama and French (2015) five-factor model.

	Value Weighted Portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber risk						
Average excess return	**0.88*****	**1.02*****	**1.13*****	**1.20*****	**1.44*****	**0.56***
	[3.20]	[3.88]	[3.73]	[4.65]	[4.19]	[1.72]
CAPM alpha	**-0.22**	**-0.06**	**-0.04**	**0.07**	**0.31**	**0.54**
	[-0.95]	[-0.42]	[-0.35]	[1.21]	[1.61]	[1.32]
FFC alpha	**-0.14**	**-0.01**	**0.03**	**0.04**	**0.24***	**0.38***
	[-1.20]	[-0.15]	[0.41]	[0.62]	[1.93]	[1.87]
FF5 alpha	**-0.16**	**-0.08**	**0.03**	**0.04**	**0.25***	**0.41****
	[-1.62]	[-0.89]	[0.39]	[0.54]	[1.89]	[2.21]
B. Characteristics						
Number of firms	615.7	615.1	615.1	615.1	615.5	-
Cyber risk	0.493	0.507	0.518	0.532	0.572	-

Table 5.3: **Average monthly excess returns and alphas (in percent)**

FFC refers to the four-factor model of Carhart (1997), and FF5 refers to the five-factor model of Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk of the portfolios. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: January 2009–December 2022

5.2.2 Before and after Florackis et al. was first released

I implement the same analysis as in Chapter 5.2.1 above, but constraining the study period to before the first release of Florackis et al. (2023) on SSRN (January 2009 until October 2020) and then after the release (November 2020 until December 2022).

Table A.4 presents the results using the period before the release. I observe that the long-short portfolio has statistically significant positive excess returns and alphas (significant at the 1% level). The outperformance of Portfolio 5 and the underperformance of Portfolio 1 is also more substantial, with the long-short portfolio having an average monthly excess return of 0.94%. Portfolio 1 has statistically significant negative alphas at the 1% level.

Table A.5 presents the results using the period after the release. Portfolio 1 has a high average monthly excess return of 2.16%, significant at the 5% level, and outperforms the other portfolios whose excess returns are not statistically significant. The long-short portfolio has negative average excess returns of -1.49%, significant at the 5% level, and statistically significant negative alphas when controlling for the market, at the 1% level. Still, the alphas are not statistically significant when controlling for the factors from Carhart (1997) or Fama and French (2015).

There could be several explanations for these results. It could be that the publication made some arbitrageurs trade stocks based on the cyber risk measure, which results in lower returns post-publication, as explained in McLean and Pontiff (2016).

It is also possible that because of cybersecurity events, high-risk firms lost value while low-risk firms appreciated. For instance, T-Mobile was a victim of a cyber attack in August 2021 during which more than 76.6 million current and former customers' information had been accessed[13]. Furthermore, the U.S Treasury Department published a report that as of June 2021, financial institutions had already reported 635 suspicious ransomware-related activities which constituted a 30% increase from all reported activity in 2020[14]. The report also found that the cost of ransomware payments increased. These events, in combination with others, could explain why Portfolio 5 has low returns and the long-short portfolio has negative returns. However, it is important to note that the study sample after the publication is much smaller, and as a result the observations could be spurious.

[13] Available at: https://www.t-mobile.com/news/network/cyberattack-against-tmobile
[14] Available at: https://cyberscoop.com/ransomware-treasury-cryptocurrency

5.3 Fama-Macbeth regressions

Table 5.4 presents the results of Fama-Macbeth regressions. Model 1 only includes the market factor. The coefficient on the market is not significant and the average adjusted R-squared is small, showing that the CAPM can not price the cyber beta-sorted portfolios. In model 2, I used the cyber risk measure and as shown, the risk premium is statistically significant and the average adjusted R-squared increases significantly. Models 3, 4 and 5 control for other common factors, and the cyber risk premium stays economically and statistically significant.

The economic interpretation of this table is that a one standard-deviation increase in cyber risk increases returns by 0.18% per month. This increase is statistically significant at the 10 or 5% level, even when controlling for other common factors.

Dependent variable: Monthly Portfolio returns					
	(1)	(2)	(3)	(4)	(5)
Market	**-0.005**		**-0.025**	**0.065**	**0.024**
	[-0.064]		[-0.429]	[0.997]	[0.275]
Cyber risk		**0.183***	**0.182****	**0.183***	**0.172****
		[1.794]	[1.994]	[1.913]	[2.037]
HML				**0.027**	**-0.012**
				[0.439]	[-0.126]
SMB				**0.069**	**0.049**
				[0.835]	[0.536]
MOM				**0.011**	
				[0.153]	
RMW					**-0.085**
					[-1.436]
CMA					**-0.088**
					[-0.816]
Constant	**1.445*****	**1.465*****	**1.457*****	**1.455*****	**1.476*****
	[5.311]	[5.540]	[5.493]	[5.413]	[5.450]
$\overline{R2_{adj}}$	0.007	0.134	0.186	0.258	0.284
MAPE	1.360	1.312	1.233	1.064	0.987

Table 5.4: **Fama-MacBeth regressions**

The betas are standardized before the second step regressions. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). MOM refers to the momentum factor from Carhart (1997). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). $\overline{R2_{adj}}$ is the average adjusted R-squared and MAPE is the mean average pricing error. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

Table A.6 presents the results of Fama-Macbeth regressions including the Fama-French 12 industries (4 of them have to be dropped due to collinearity). The cyber risk premium is reduced to 0.156% but is still significant at the 10% level.

5.4 GRS test

I implement the GRS test as follows: I build 20 value-weighted portfolios sorted on cyber risk, then compute the GRS test statistic using the five-factor model from Fama and French (2015) and the same model plus the cyber risk factor. I report the results then I repeat this procedure, sorting the portfolios on market beta, firm size, and book-to-market ratio.

Table 5.5 presents the results. I observe that the GRS test statistic is smaller for the model containing the cyber risk factor when sorting on cyber risk, size and book-to-market. Interestingly, when sorting on firm size and book-to-market, I can reject the null hypothesis that $\alpha_i = 0 \ \forall i$, for the five-factor model but not for the model containing the cyber risk factor. It is not the case when sorting on market beta, however, I can not reject the null hypothesis for either model. In fact, I can not reject this hypothesis for most tests reported in this table.

	GRS	p value	$\overline{R2}$	GRS	p value	$\overline{R2}$
	Sorted on cyber risk			Sorted on market beta		
FF5	1.211	0.253	0.869	0.712	0.802	0.783
FF5 + CyberFactor	0.947	0.530	0.886	0.825	0.680	0.801
	Sorted on size			Sorted on book-to-market		
FF5	1.490	0.093	0.879	1.709	0.038	0.878
FF5 + CyberFactor	1.458	0.106	0.880	1.417	0.124	0.883

Table 5.5: **GRS test statistics**

R squared values are averaged over the 20 portfolios. FF5 refers to the five-factor model from Fama and French (2015) and CyberFactor refers to the long-short portfolio from Chapter 5.2.1.

These results suggest that a subset of factors could explain the cross-section of returns.

5.5 Bayesian factor model selection

Given the results of the GRS tests, a subset of factors could potentially explain the cross-section of returns. The analysis explained in Chapter 4.2.3 allows us to determine the combination of factors that is best in terms of pricing returns. Figure 5.4 presents the posterior probabilities of the 5 most likely models, ranked at the end of the sample. All five models contain the cyber risk factor and the model with the highest probability, of 21.18%, is the model containing the Market, book-to-market, investment, operating profitability, and cyber risk factors.

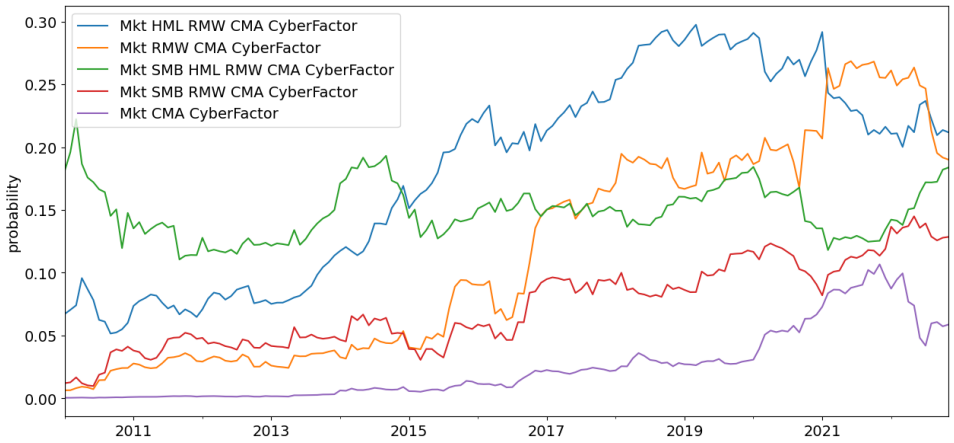

Figure 5.4: **Factor model posterior probabilities**

The figure shows the posterior probabilities for the top 5 models, ranked at the end of the sample. Mkt refers to the excess return of the market from the Kenneth French data repository. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). CyberFactor refers to the long-short portfolio from Chapter 5.2.1. Prior Multiple = 1.5

Figure 5.5 presents the cumulative factor probabilities, that is the sum of probabilities of all models containing the factor. The cyber risk factor has a cumulative probability of 91.66% at the end of the sample. The investment and operating profitability factors also have very high cumulative probabilities, unlike the remaining factors.

Finally, I study the sensitivity of the model probabilities to the prior multiple. I repeat the analysis using three other values of prior multiple: 1.25, 2 and 3 (similarly to Barillas and Shanken (2018)). Table 5.6 reports the posterior model probabilities at the end of the sample for the top five models for each prior multiple. I observe that the top five models are the same for each prior multiple. Furthermore, the book-to-market factor is no longer in the most likely model for higher values of the prior multiple.

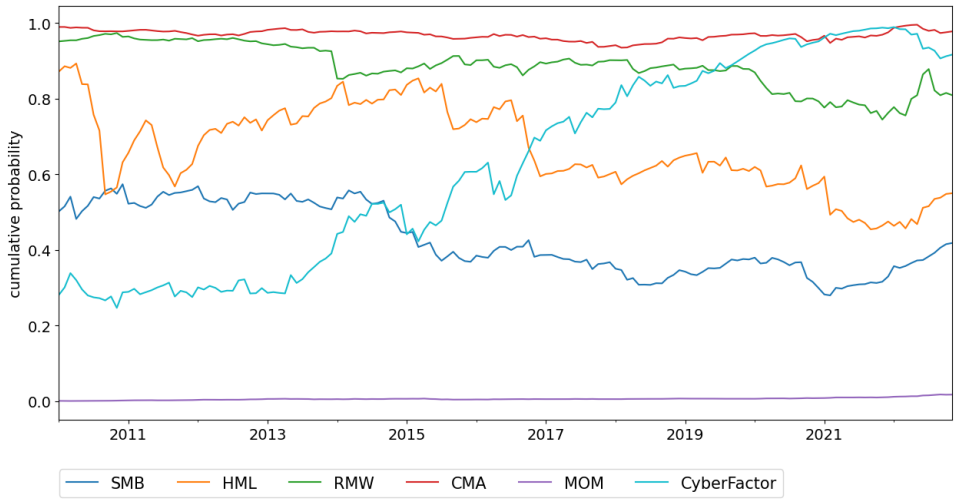

Figure 5.5: **Cumulative posterior factor probabilities**

Cumulative posterior probabilities are the sum of probabilities of all models containing the factor. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). MOM refers to the momentum factor from Carhart (1997). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). CyberFactor refers to the long-short portfolio from Chapter 5.2.1. Prior Multiple = 1.5

5.6 Robustness tests

5.6.1 Long-run cyber risk

By design, the cyber risk computed in Chapter 4.1.4 depends only on the most recent 10-K statement of each company. It could be possible that a firm discusses cybersecurity concerns and risks extensively in its 10-K statement in year T, for example, because of an increasing number of cyberattacks in the industry, resulting in a high cyber risk score. Having focused on cyber risk in year T and not having been attacked itself, the firm could decide not to talk about cyber risks in its next 10-K statements (in years $T + i$), or not as much as in year T, even though it still has similar cyber risks. These cases would be missed by the previously computed cyber risk measure as it has no memory, and I would be measuring shocks to cyber risk.

Using the cyber risk defined in Chapter 4.1.4, I compute the expanding average cyber risk score in order to study the long-run cyber risk of firms. That is the long-run cyber risk score in year T is the average of its simple cyber risk scores from year 2008 to year T. This new measure could account for the companies described above. The advantage of using the long-run average instead of the long-run maximum is that it does

Prior Multiple	1.25	1.5	2	3
Mkt HML RMW CMA CyberFactor	19.40	21.18	21.31	18.79
Mkt RMW CMA CyberFactor	16.54	19.01	21.89	26.08
Mkt SMB HML RMW CMA	18.38	18.38	15.92	10.41
Mkt SMB RMW CMA CyberFactor	11.92	12.85	12.81	11.24
Mkt CMA CyberFactor	5.76	5.87	7.15	11.18

Table 5.6: **Prior sensitivity of the posterior model probabilities**

The table shows the posterior model probabilities (in percent) for the top 5 models (ranked at the end of the sample) for different values of the prior multiple. The top 5 models are the same for every prior multiple. Mkt refers to the excess return of the market from the Kenneth French data repository. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). CyberFactor refers to the long-short portfolio from Chapter 5.2.1

not discard the observations. This is beneficial when considering firms in the following situation: consider a firm that discusses cyber risks in its 10-K statement in year T, following a cyberattack or data breach. The firm might purchase protection (insurance, software,...) and hence minimize its future cyber risk, and not discuss this risk in its future 10-K statements. The low cyber risk scores in the upcoming years are representative of the reality of the firm and should not be discarded.

I repeat the portfolio sorts from Chapter 5.2 using the long-run cyber risk to sort firms. Table A.7 presents the results. I observe no significant change from Table 5.3 and the long-short portfolio remains significant at the 5% or 10% level. These results could be an indication that firms incorporate all available information in their newest 10-K statements regarding their cyber risk and hence incorporating information from past statements does not improve the estimation of the cyber risk.

5.6.2 Controlling for cybersecurity firms

The cyber risk score constructed in Chapter 4.1.4 does not make a distinction between firms that discuss cybersecurity because they consider it a risk and firms that are cybersecurity solutions providers, for example, Fortinet[15].

As there is no dedicated cybersecurity industry classification, I identify cybersecurity firms using the HACK ETF[16]. As explained in the fund's description, this ETF invests in companies providing cybersecurity solutions that include hardware, software, and services. As cyber-

[15]Available at: https://www.fortinet.com/

[16]Available at: https://etfmg.com/funds/hack/

security providers, these firms are expected to discuss cybersecurity in their 10-K statements extensively, resulting in a false high cyber risk score. Indeed, I observe that these firms have an average score of 0.59, which is in the top 3% of cyber risk scores. I repeat the analysis from Chapter 5.2 and I exclude the holdings of the HACK ETF from the universe of firms.

Table A.8 shows the results. The results for Portfolios 1, 2, and 3 are unchanged, and the excess returns and alphas of Portfolios 4 and 5 increase. Furthermore, I observe that the t-statistics on Portfolios 4 and 5 increase as well. These results support the view that it is the cyber risk that is priced.

Chapter 6

Conclusion

In this thesis, I implement a doc2vec model to estimate firms' cyber risk based on their 10-K statements. I then use this cyber risk measure in various asset pricing tests. The results support the view that cyber risk is priced in the cross-section of firms. Indeed, a long-short strategy on cyber risk sorted portfolios has a positive and statistically significant alpha with respect to traditional factor models and an average monthly excess return of 0.56%. I also perform this analysis limiting to the period before and then after the first release of Florackis et al. (2023) on SSRN. I find that while the long-short strategy has statistically significant excess return and alphas at the 1% level before the release, it has negative average excess returns after. Furthermore, using Fama-Macbeth regressions, I show that cyber risk has a significant risk premium. Using the GRS test of Gibbons et al. (1989) and the methodology from Barillas and Shanken (2018), I show that the cyber risk factor helps price stocks and is present in the five most likely factor models. I also compute the long-run cyber risk as the expanding average of the cyber risk. I perform the portfolio sorts and observe that the results are very similar. Hence, I conclude that firms incorporate all available information about cyber risk in their newest 10-K statement. Finally, I exclude cybersecurity firms from the sample and perform the portfolio sorts. I find that the average monthly returns of the two high cyber risk portfolio increase, and the others are left unchanged, supporting the view that the alpha is due to the cyber risk.

Nevertheless, this paper has some limitations. First, using disclosures implicitly relies on the firms' willingness to disclose information about their cyber risks. Even though there are no mandatory cybersecurity disclosure rules (but there might be soon, see the SEC's press re-

lease[17]), the SEC has had cybersecurity guidance since 2011[18], meaning most firms already discuss cyber risks in their 10-K statements. Without a mandatory cyber risk section, however, the length and the location of the cybersecurity information in the 10-K statements are inconsistent from one statement to another. This issue is limited by the methodology used in this paper. Indeed, I use the complete text from the 10-K statements; hence even if the location of the cybersecurity texts is inconsistent, I can identify them using the doc2vec methodology.

Next, the model itself can not distinguish between firms discussing cybersecurity because they consider it a risk and firms that are cybersecurity solutions providers. However, the number of listed firms that are purely cybersecurity providers is minimal, and as shown in chapter 5.6.2, excluding such firms only improves the results.

This research can be extended in two directions. First, the study can be repeated for stock markets of different countries. Second, it is possible to estimate firms' exposure to other risk factors using this machine learning approach, for instance, legislative risk. These risks could be readily captured, by changing the MITRE knowledgebase to one that is close to the risk under scrutiny.

[17]Available at: https://www.sec.gov/news/press-release/2023-52

[18]Available at: https://www.sec.gov/divisions/corpfin/guidance/cfguidance-topic2.htm

Bibliography

Adosoglou, G., Lombardo, G., Pardalos, P. M., 2021. Neural network embeddings on corporate annual filings for portfolio selection. Expert Systems with Applications 164, 114053.

Anderson, R., Barton, C., Boehme, R., Clayton, R., Ganan, C., Grasso, T., Levi, M., Moore, T., Vasek, M., 2019. Measuring the changing cost of cybercrime. Workshop on the Economics of Information Security 18, 1–32.

Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M. J. G., Levi, M., Moore, T., Savage, S., 2013. Measuring the cost of cybercrime. Workshop on the Economics of Information Security 11, 265–300.

Andreadis, L., Kalotychou, E., Louca, C., Lundblad, C. T., Makridis, C., 2023. Cyberattacks, media coverage and municipal finance. Available at https://dx.doi.org/10.2139/ssrn.4473545

Barillas, F., Shanken, J., 2017. Which alpha? Review of Financial Studies 30, 1316–1338.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. Journal of Finance 73, 715–754.

Bouveret, A., 2018. Cyber risk for the financial sector: A framework for quantitative assessment. Available at http://dx.doi.org/10.2139/ssrn.3203026

Campbell, K., Gordon, L. A., Loeb, M. P., Zhou, L., 2003. The economic cost of publicly announced information security breaches: Empirical evidence from the stock market. Journal of Cybersecurity 11, 431–448.

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

Cochrane, J. H., 2005. Asset pricing. Princeton University Press.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. The Journal of Finance 47, 427–465.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

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

Farrow, S., Szanton, J., 2016. Cybersecurity investment guidance: Extensions of the Gordon and Loeb Model. Journal of Information Security 7, 15–28.

Florackis, C., Louca, C., Michaely, R., Weber, M., 2023. Cybersecurity risk. Review of Financial Studies 36, 351–407.

Gibbons, M. R., Ross, S. A., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–1152.

Gordon, L. A., Loeb, M. P., 2002. The economics of information security investment. ACM Transactions on Information and System Security 5, 438–457.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., Zhou, L., 2015a. Externalities and the magnitude of cyber security underinvestment by private sector firms: A modification of the Gordon-Loeb Model. Journal of Information Security 6, 24–30.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., Zhou, L., 2015b. The impact of information sharing on cybersecurity underinvestment: A real options perspective. Journal of Accounting and Public Policy 34, 509–519.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., Zhou, L., 2015c. Increasing cybersecurity investments in private sector firms. Journal of Cybersecurity 1, 3–17.

Gordon, L. A., Loeb, M. P., Sohail, T., 2010. Market value of voluntary disclosures concerning information security. Management Information Systems Quarterly 34, 567–594.

Gordon, L. A., Loeb, M. P., Zhou, L., 2011. The impact of information security breaches: Has there been a downward shift in costs? Journal of Computer Security 19, 33–56.

Harvey, C. R., Liu, Y., Zhu, H., 2016. ...and the cross-section of expected returns. Review of Financial Studies 29, 5–68.

Hausken, K., 2006. Returns to information security investment: The effect of alternative information security breach functions on optimal investment and sensitivity to vulnerability. Information Systems Frontiers 8, 338–349.

Hilary, G., Segal, B., Zhang, M. H., 2016. Cyber-risk disclosure: Who cares? Available at http://dx.doi.org/10.2139/ssrn.2852519

Jamilov, R., Rey, H., Tahoun, A., 2021. The anatomy of cyber risk. Available at https://doi.org/10.3386/w28906

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48, 65–91.

Jensen, J., Paine, F., 2023. Municipal cyber risk. Available at https://weis2023.econinfosec.org/wp-content/uploads/sites/11/2023/06/weis23-jensen.pdf

Johnson, M., Kang, M. J., Lawson, T., 2017. Stock price reaction to data breaches. Journal of Finance Issues 16, 1–13.

Kamiya, S., Jun-Koo, K., Jungmin, K., Milidonis, A., Stulz, R. M., 2021. Risk management, firm reputation, and the impact of successful cyber-attacks on target firms. Journal of Financial Economics 139, 719–749.

Lau, J. H., Baldwin, T., 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*, Association for Computational Linguistics, Berlin, Germany, pp. 78–86.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Xing, E. P., Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, PMLR, Bejing, China, pp. 1188–1196.

Lelarge, M., 2012. Coordination in network security games: A monotone comparative statics approach. IEEE Journal on Selected Areas in Communications 30, 2210–2219.

Lending, C., Minnick, K., Schorno, P. J., 2018. Corporate governance, social responsibility, and data breaches. Financial Review 53, 413–455.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics 47, 13–37.

McLean, R. D., Pontiff, J., 2016. Does academic research destroy stock return predictability? Journal of Finance 71, 5–32.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space.

Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica 34, 768–783.

Newey, W. K., West, K. D., 1994. Automatic lag selection in covariance matrix estimation. Review of Economic Studies 61, 631–653.

Romanosky, S., 2016. Examining the costs and causes of cyber incidents. Journal of Cybersecurity 2, 121–135.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance 19, 425–442.

Tosun, O. K., 2021. Cyber-attacks and stock market activity. International Review of Financial Analysis 76, 1–15.

Wang, S. S., 2019. Integrated framework for information security investment and cyber insurance. Pacific-Basin Finance Journal 57, 1–12.

Willemson, J., 2006. On the Gordon and Loeb Model for information security investment. Workshop on Economics and Information Security 5, 1–12.

Appendix

Method	Vector Size	Window Size	Min Count	Sub-Sampling	Negative Sampling	Epoch
DBOW	300	15	5	10^{-5}	5	20

Table A.1: **Baseline doc2vec parameters**

The parameters of the baseline model are taken from Lau and Baldwin (2016). DBOW stands for distributed bag-of-words.

Score	Preprocessed paragraph	Ticker	Tactic
0.593	currently available internet browsers allow users modify browser settings remove cookies prevent cookies stored hard drives however third persons able penetrate network security gain access otherwise misappropriate users personal information subject liability liability include claims misuses personal information unauthorized marketing purposes unauthorized use credit cards	VSTY	Defense Evasion
0.590	network security data recovery measures may adequate protect computer viruses break ins similar disruptions unauthorized tampering computer systems theft sabotage type security breach respect proprietary confidential information electronically stored including research clinical data material adverse impact business operating results financial condition	LXRX	Collection
0.583	domain names derive value individual ability remember names therefore assurance domain name lose value example users begin rely mechanisms domain names access online resources government regulation internet regulation increasing number laws regulations pertaining internet	VSTY	Credential Access
0.577	perceived actual unauthorized disclosure information collect breach security harm business factors beyond control cause interruptions operations may adversely affect reputation marketplace business financial condition results operations timely development implementation continuous uninterrupted performance hardware network applications internet systems including may provided third parties important facets delivery products services customers	MDAS	Credential Access
0.571	unauthorized parties may attempt copy aspects products obtain use information regard proprietary others may independently develop otherwise acquire similar competing technologies methods design around patents cases rely trade secret laws confidentiality agreements protect confidential proprietary information processes technology	CSCD	Collection

Table A.2: **Top scoring paragraphs from the doc2vec validation sample**

The paragraphs are shown after preprocessing (as described in section 4.1.1). Tactic refers to the MITRE tactic the paragraph is most similar to, as measured by cosine similarity. The tickers of the 10 companies in the validation sample are CSCD, GTS, LXRX, MDAS, PBY, PZZA, UMH, VALU, VSTY and VXRT.

Score	Preprocessed paragraph	Ticker	Tactic
0.570	possible cookies may become subject laws limiting prohibiting use term cookies refers information keyed specific server file pathway directory location stored user hard drive possibly without user knowledge used among things track demographic information target advertising	VSTY	Discovery
0.561	cannot certain advances computer capabilities discoveries field cryptography developments result compromise breach algorithms use protect content transactions website proprietary information databases anyone able circumvent security measures misappropriate proprietary confidential customer company information cause interruptions operations	VSTY	Impact
0.558	ordering delivery customers ready place order proceed shopping cart function directly checkout page orders placed online website via toll free telephone number customer service agents available take orders customers access internet uncomfortable placing order online	VSTY	Credential Access
0.557	process allows identify catalogue embryonic stem cell clone dna sequence trapped gene select embryonic stem cell clones dna sequence generation knockout mice used gene trapping technology automated process create omnibank library frozen gene knockout embryonic stem cell clones identified dna sequence relational database	LXRX	Persistence
0.556	believe systematic biology driven approach technology platform makes possible provide substantial advantages alternative approaches drug target discovery particular believe comprehensive nature approach allows uncover potential drug targets within context mammalian physiology might missed narrowly focused efforts	LXRX	Discovery
0.554	concerns security internet may reduce use website impede growth significant barrier confidential communications internet need security rely ssl encryption technology designed prevent customer credit card data transaction process current credit card practices merchant liable fraudulent credit card transactions case transactions process merchant obtain cardholder signature	VSTY	Credential Access

Table A.2: **Top scoring paragraphs from the doc2vec validation sample (continued)**

Variable	Description	Source
Firm size (ln)	ln(total assets [at])	Compustat
Firm Age (ln)	ln(years) since the firm first appeared in Compustat	Compustat
Book to market ratio	Common equity [ceq] / market equity [prc*shrout]	Compustat and CRSP
Tobin's Q	(Total assets - common equity + market equity) / total assets	Compustat and CRSP
ROA	Net income [ni] / total assets	Compustat
Market Beta	5-year rolling market beta [beta]	Compustat
Intangible/Assets	Intangible assets [intan] / total assets	Compustat
Debt/assets	Total Debt / Total Assets [debt_assets]	WRDS Financial Ratios
ROE	Net Income / Book Equity [roe]	WRDS Financial Ratios
Price/Earnings	Stock Price / Earnings [pe_exi]	WRDS Financial Ratios
Profit Margin	Gross Profit / Sales [gpm]	WRDS Financial Ratios
Asset Turnover	Sales / Total Assets [at_turn]	WRDS Financial Ratios
Cash Ratio	(Cash + Short-term Investments) / Current Liabilities [cash_ratio]	WRDS Financial Ratios
Sales/Invested Capital	Sales per dollar of Invested Capital [sale_invcap]	WRDS Financial Ratios
Capitalization Ratio	Long-term Debt / (Long-term Debt + Equity) [capital_ratio]	WRDS Financial Ratios
R&D/Sales	R&D expenses / Sales [RD_SALE]	WRDS Financial Ratios
ROCE	Earnings Before Interest and Taxes / average Capital Employed [roce]	WRDS Financial Ratios

Table A.3: **Variable definitions**

The names of the variables as found on CRSP and Compustat are in brackets.

	Value Weighted Portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber risk						
Average excess return	**0.70****	**0.97*****	**1.10*****	**1.23*****	**1.64*****	**0.94*****
	[2.41]	[3.58]	[3.69]	[5.76]	[5.76]	[3.23]
CAPM alpha	**-0.52*****	**-0.20**	**-0.14**	**0.06**	**0.51****	**1.03*****
	[-3.11]	[-1.50]	[-1.07]	[0.95]	[2.40]	[2.95]
FFC alpha	**-0.28*****	**-0.06**	**0.00**	**-0.03**	**0.27***	**0.55*****
	[-3.58]	[-0.73]	[0.01]	[-0.52]	[1.91]	[2.83]
FF5 alpha	**-0.26*****	**-0.08**	**0.03**	**-0.06**	**0.30***	**0.56*****
	[-3.45]	[-0.88]	[0.36]	[-0.95]	[1.90]	[3.01]
B. Characteristics						
Number of firms	600.4	599.9	599.9	599.9	600.2	-
Cyber risk	0.490	0.504	0.515	0.529	0.570	-

Table A.4: **Average monthly excess returns and alphas (in percent) before the first release of Florackis et al. on SSRN**

FFC refers to the four-factor model from Carhart (1997) and FF5 refers to the five-factor model from Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk of the portfolios. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: January 2009–October 2020 (before the first release of Florackis et al. (2023) on SSRN)

	Value Weighted Portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber risk						
Average excess return	**2.16****	**1.56**	**1.34**	**1.55**	**0.67**	**-1.49****
	[2.18]	[1.64]	[0.97]	[1.19]	[0.44]	[-2.11]
CAPM alpha	**1.34*****	**0.71*****	**0.34**	**0.54*****	**-0.40***	**-1.74*****
	[4.37]	[2.91]	[1.15]	[3.44]	[-1.67]	[-3.35]
FFC alpha	**0.57***	**0.19**	**-0.07**	**0.54*****	**0.17**	**-0.39**
	[1.76]	[0.71]	[-0.37]	[3.10]	[0.61]	[-0.67]
FF5 alpha	**0.34**	**-0.15**	**-0.15**	**0.65*****	**0.11**	**-0.23**
	[0.89]	[-0.53]	[-0.74]	[3.63]	[0.38]	[-0.34]
B. Characteristics						
Number of firms	695.5	695.0	694.9	695.0	695.2	-
Cyber risk	0.504	0.523	0.536	0.550	0.582	-

Table A.5: **Average monthly excess returns and alphas (in percent) after the first release of Florackis et al. on SSRN**

FFC refers to the four-factor model from Carhart (1997) and FF5 refers to the five-factor model from Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk of the portfolios. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: November 2020–December 2022 (after the first release of Florackis et al. (2023) on SSRN)

Dependent variable: Monthly Portfolio returns	
	(6)
Cyber risk	**0.156***
	[1.951]
HML	**-0.116**
	[-1.236]
SMB	**0.200**
	[0.951]
RMW	**0.062**
	[0.753]
CMA	**-0.173**
	[-0.925]
Consumer Durables	**-0.013**
	[-0.212]
Manufacturing	**-0.094**
	[-1.261]
Energy	**0.059**
	[0.645]
Chemicals	**0.062**
	[0.909]
Telecommunications	**-0.162****
	[-2.19]
Retail	**0.139**
	[1.272]
Healthcare	**-0.003**
	[-0.046]
Finance	**-0.019**
	[-0.211]
Constant	**1.329*****
	[4.081]
$\overline{R2_{adj}}$	0.307
MAPE	0.636

Table A.6: **Fama-MacBeth regressions with industries**

The betas are standardized before the second step regressions. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). The industries correspond to the 12 Fama-French industries, obtained from the Kenneth French data repository. The "Business Equipment", "Consumer NonDurables", "Other",and "Utilities" industries are dropped due to high colinearity with the other industries and factors (as measured by the Variance Inflation Factor). $\overline{R2_{adj}}$ is the average adjusted R-squared and MAPE is the mean average pricing error. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

	Value Weighted Portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by the long-run cyber risk						
Average excess return	**0.86*****	**1.13*****	**1.14*****	**1.18*****	**1.45*****	**0.60***
	[3.00]	[3.94]	[3.93]	[4.35]	[4.23]	[1.70]
CAPM alpha	**-0.26**	**-0.01**	**-0.03**	**0.11***	**0.31**	**0.57**
	[-1.05]	[-0.04]	[-0.27]	[1.71]	[1.52]	[1.33]
FFC alpha	**-0.17**	**0.007**	**0.05**	**0.10**	**0.23***	**0.40****
	[-1.56]	[0.67]	[0.75]	[1.24]	[1.83]	[2.01]
FF5 alpha	**-0.17***	**0.03**	**-0.01**	**0.06**	**0.25***	**0.43****
	[-1.80]	[0.34]	[-0.16]	[0.93]	[1.94]	[2.28]
B. Characteristics						
Number of firms	615.7	615.1	615.1	615.1	615.5	-
Long-run cyber risk	0.490	0.501	0.511	0.524	0.567	-

Table A.7: **Average monthly excess returns and alphas (in percent) using the long-run cyber risk**

The portfolios are sorted using the long-run cyber risk. FFC refers to the four-factor model from Carhart (1997) and FF5 refers to the five-factor model from Fama and French (2015). Panel B shows the average number of firms in each portfolio and the portfolio's average long-run cyber risk. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: January 2009-December 2022

	Value Weighted Portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber risk						
Average excess return	**0.88*****	**1.02*****	**1.13*****	**1.22*****	**1.45*****	**0.57***
	[3.19]	[3.86]	[3.71]	[4.73]	[4.18]	[1.74]
CAPM alpha	**-0.22**	**-0.07**	**-0.05**	**0.08**	**0.33**	**0.55**
	[-0.94]	[-0.44]	[-0.39]	[1.38]	[1.67]	[1.34]
FFC alpha	**-0.14**	**-0.02**	**0.03**	**0.05**	**0.25****	**0.39***
	[-1.18]	[-0.19]	[0.34]	[0.81]	[2.027]	[1.91]
FF5 alpha	**-0.16**	**-0.08**	**0.03**	**0.06**	**0.26****	**0.41****
	[-1.61]	[-0.93]	[0.35]	[0.72]	[1.97]	[2.26]
B. Characteristics						
Number of firms	611.9	611.3	611.3	612.3	611.7	-
Cyber risk	0.493	0.507	0.518	0.532	0.571	-

Table A.8: **Average monthly excess returns and alphas (in percent), cyber firms dropped**

Cybersecurity firms are dropped from the sample. FFC refers to the four-factor model from Carhart (1997), and FF5 refers to the five-factor model from Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk of the portfolios. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: January 2009-December 2022