Cyber risk and the cross-section of stock returns

Daniel Celeny* and Loïc Maréchal†

February 8, 2024

Abstract

We extract firms' cyber risk with a machine learning algorithm measuring the proximity between their disclosures and a dedicated cyber corpus. Our approach outperforms dictionary methods, uses full disclosure and not devoted-only sections, and generates a cyber risk measure uncorrelated with other firms' characteristics. We 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, we show that our cyber risk factor is the essential feature that allows any multi-factor model to price the cross-section of stock returns.

JEL classification: C45, C58, G12.

Keywords: natural language processing, machine learning, asset pricing.

*Swiss Finance Institute, EPFL - Cyber-Defence Campus, armasuisse S+T. daniel.celeny@alumni.epfl.ch
†HEC Lausanne, University of Lausanne, loic.marechal@unil.ch

This document results from a research project funded by the Cyber-Defence Campus, armasuisse Science and Technology. A preprint of this research is available on SSRN. We appreciate helpful comments from seminar participants at the Cyber Alp Retreat 2023. We also thank Pierre Collin-Dufresne for his invaluable comments. Our code will be available at `https://github.com/technometrics-lab/17-Cyber-risk_and_the_cross-section_of_stock_returns`
Corresponding author: Daniel Celeny e-mail: daniel.celeny@alumni.epfl.ch, Cyber-Defence Campus, Innovation Park, EPFL, 1015 Lausanne.

1. Introduction

Due to the increasing digitalization of our environment, usage of Internet-of-Things devices, and geopolitical interests, the number of cyberattacks and costs constantly increase. 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 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. In a recent interview[1], Mario Greco, CEO of Zurich Insurance Group, said that cyberattacks 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".

Florackis, Louca, Michaely, and Weber (2023) and Jamilov, Rey, and Tahoun (2021) use dictionary methods for cyber risk extractions in 10-K filings and earning calls, respectively.[2] We argue that this approach is unsuitable for this purpose.

In this paper, we 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 that shows up in their stock returns. To do this, we collect financial fillings, monthly returns, and other firm characteristics for over 7,000 firms listed on US stock markets between January 2007 and December 2022. We use a machine learning algorithm, the "Paragraph Vector", in combination with the MITRE ATT&CK cybersecurity knowledgebase to score each firm's filing based on its cybersecurity content.

We find evidence that our cyber risk does not correlate with firm size, book-to-market, beta, and other standard firms' characteristics known to help price stock returns. At the aggregated level, our 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 exceptionally narrow (standard deviation of 0.03). We compare our cyber risk measure across Fama-French 12 industries and find results supporting our intuition, with "Business Equipment" and "Telephone and Television Transmission" being the riskiest and "Oil and Gas" and "Utilities", the safest.

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

[2] See also, Jiang, Khanna, Yang, and Zhou, 2023; Liu, Marsh, and Xiao, 2022

We 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 particularly well before the first mention of a cyber risk premium on SSRN in November 2020 by Florackis et al. (2023), with an average annual excess return of 11.88%, and is statistically significant at the 1% level. Double sorts confirm that cyber risk captures a variation in stock returns when controlling for other factors.

We use asset pricing tests and find that the cyber risk exposure 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), we show that the optimal subset of factors pricing stock returns always includes our cyber risk factor.

We conduct tests to verify the robustness of our factor. First, we control that our baseline measure, revised at each new filing, captures the latent cyber risk and not the immediate effect of a cyberattack. To do so, we build a long-run cyber risk measure capturing the cumulative cyber risk effect. Our results are virtually unchanged. Second, we control for the possibility that firms in the cybersecurity business have risk occurrences that might positively affect them. We also do not find any differences after that control.

Finally, we compare our model to the dictionary approach used by Florackis et al. (2023). While the two measures are positively correlated, our measure performs better, especially for firms that were assigned zero cyber risk using the dictionary approach.

The remainder of the paper proceeds as follows. Section 2 introduces the existing literature and develops related hypotheses. Section 3 presents the data and methods, Section 4 details the results, and Section 5 concludes.

2. Literature review

2.1. 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. For example, each piece of text is represented by a dense vector that can be used for text classification and sentiment analysis. 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 for neural network embeddings, such as the ones obtained using the "Paragraph Vector" algorithm.

2.2. Cybersecurity costs

2.2.1. Direct estimations

Anderson, Barton, Böhme, Clayton, van Eeten, Levi, Moore, and Savage (2013) perform a systematic study of the costs of cybercrime. They differentiate direct, indirect, and defense costs and disentangle the 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 defense costs, however, are much higher for transitional and new crimes. They conclude that we should spend less anticipating cybercrime and more in response.

Anderson, Barton, Boehme, Clayton, Ganan, Grasso, Levi, Moore, and Vasek (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 citizens have fallen. They conclude that economic optimality would be spending less on cyberattack prevention and more on response and law enforcement. Bouveret (2018) documents cyber risk worldwide in the financial sector by analyzing the different types of cyber incidents and identifying patterns. He uses a Value at Risk (VaR) type of framework. He finds an average loss due to cyberattacks of USD 97 bn at the country level and a VaR between USD 147 and 201 bn. He concludes that there are sizeable potential aggregated losses in the financial sector, several orders of magnitude higher than the cyber insurance market can cover. 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 USD 6 mln and a median cost of USD 170,000 (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.

2.2.2. Indirect estimations with municipal bonds

Andreadis, Kalotychou, Louca, Lundblad, and Makridis (2023) study the impact of information dissemination about cyberattacks through significant news sources on municipalities' access to finance, focusing on the municipal bond market. They employ a difference-in-differences framework and find that the cumulative number of cyberattacks covered by county-level news articles and the corresponding number of county-level cyberattack news articles significantly adversely affect 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 declined and IT spending increased. They argue that the declining bond yields are driven by a decrease in the town's cyber risk due to increased IT spending.

2.2.3. Indirect estimations with stock price reaction

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. The findings from the study suggest that in recent years, average information security breaches have become less costly, and there seems to be a shift in attitude among investors toward viewing information security breaches as creating 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 include confidential information. Johnson, Kang, and Lawson (2017) also study cumulative abnormal returns around cyber security events. They show that, on average, publicly traded firms in the U.S. lost 0.37% of their equity value when a data breach occurs. Breaches resulting from payment card fraud contribute more to negative announcement returns than the other breach types, and the adverse effects are more important for firms with card breaches larger than the average. For the average firm, these breaches result in a 3% decline in firm equity value.

Lending, Minnick, and Schorno (2018) study the relationship between corporate governance, social responsibility, and the probability of data breaches, and they measure the changes in stock returns following data breaches. They 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 selling 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 adverse 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.

2.2.4. 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 about information security could mitigate potential litigation costs and lower the firm's cost of capital by reducing the information asymmetry between a firm's management and its investors. 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 et al. (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. Jamilov et al. (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: the long-short portfolio built on cybersecurity beta sorted portfolios has an average annual return of -3.3% (the negative sign is because the factor captures shocks to cybersecurity).

We are only aware of four studies focusing on cyber risk and its factor structure using disclosures). However, these studies use a dictionary 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 consider the context of the keywords, only their presence in the disclosures. We use the "Paragraph Vector" algorithm to fill this gap to build a cyber risk measure using firms' 10-K statements. Hence, we define our 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

3. Data and methodology

3.1. Market data

We download public equity data from Wharton Research Data Services (WRDS), using the data from the Center for Research in Security Prices (CRSP) and S&P Global Market Intelligence's Compustat database. We report the list of variables in Table A3. We write 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 so that all of the retained firms have at least one 10-K statement available. We extract monthly stock returns and financial ratios for 7,059 firms between January 2007 and December 2022. Figure 1 depicts the industry distribution of these firms using the Fama-French 12 industry classification.

We 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[3].

[Insert Figure 1 here.]

[3]Available at: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

3.2. 10-K statements

10-K statements are financial filings publicly traded companies submit annually to the US Securities and Exchange Commission (SEC). They contain information such as companies' financial statements, risk factors, and executive compensation. We use 10-K statements to build a cyber risk measure (detailed in Section 3.4).

To download these statements, we use the index files from the SEC's Edgar archives[4]. 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, we use their Central Index Key (CIK), a number used by the SEC to identify corporations and individuals who have filed disclosures. We develop a Python script that goes through these index files and identifies URLs that correspond to 10-K statements using the Form Type entry. Using the CIK entry, these URLs are then matched to one of the 7,059 firms mentioned in Section 3.1. We identify 60,470 10-K statements, corresponding to 8.6 statements per firm on average. Figure 2 shows the number of 10-Ks filed per year. This number increases from 3,301 in 2007 to 5,370 in 2022.

[Insert Figure 2 here.]

3.3. Cybersecurity tactics

We use the MITRE ATT&CK[5] 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 illustrates the structure of the knowledge base. Each sub-technique has a short description describing it. Table 1 shows two sub-technique descriptions from the knowledge base.

[Insert Figure 3 here.]

[Insert Table 1 here.]

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 used in Section 3.4.

[4]Available at: https://www.sec.gov/Archives/edgar/full-index/
[5]Available at: https://attack.mitre.org/

3.4. Methodology

3.4.1. Text preprocessing

10-K statements can be downloaded from the SEC Archives as HTML files (as explained in Section 3.2). We use the BeautifulSoup[6] Python library to extract the usable text from these files. We remove the punctuation and numbers and set all letters to lowercase. Given the resulting texts, we develop a Python script that uses the wordfreq[7] and NLTK[8] 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 essential cybersecurity-related words.

After pre-processing, the average length of the cybersecurity sub-technique descriptions from MITRE ATT&CK is close to 40 words ($\approx$ 39.7). Based on this number, we write a Python algorithm to merge consecutive sentences from 10-K statements into paragraphs with an average length of close to 40 words after pre-processing. On average, we 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.

3.4.2. 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 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.

The model has two versions: 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 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 illustrates the two models.

[Insert Figure 4 here.]

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

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

We use the implementation by Gensim called doc2vec[9]. To train the model, we 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. We train DM and DBOW doc2vec models with various vector dimensions, epochs, and window sizes using this sample. The baseline for the hyperparameters is taken from Lau and Baldwin (2016) (see table A1).

To choose the best model, we 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). We choose the best model where the proportion of the highest-scoring cyber risk-related paragraphs is the highest. Table 2 presents the parameters of the best-performing doc2vec model, which is used for the remainder of this study. Table A2 gives the top-scoring paragraphs from the validation sample (after pre-processing).

[Insert Table 2 here.]

3.4.3. Cosine similarity

We use cosine similarity to measure the distance between the vector representations of two paragraphs, *i.e.*, 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 Section 3.4.4, we 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 $sim = max(0, \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|})$.

3.4.4. Cyber risk score

The cyber risk score is based on the cosine similarities with the cybersecurity descriptions from MITRE ATT&CK. First, we compute the vector representation of every paragraph of every 10-K statement using the trained doc2vec model. We also compute the vector representation of every sub-technique description from MITRE ATT&CK. Next, we 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,

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

we compute the score of a 10-K statement as the average score of the 1% of its highest-scoring paragraphs.

This algorithm assumes that a typical 10-K statement has at most six or seven cyber risk-related paragraphs, representing 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 solely dependent on paragraphs that are most likely to be cyber risk-related.

As Section 3.4.3 mentions, the cosine similarity takes values between -1 and 1. We only consider 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, we 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 those obtained using dictionary methods such as in Jamilov et al. (2021).

Figure 5 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. We 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.

[Insert Figure 5 here.]

3.5. Asset pricing tests

We use the two-step Fama-MacBeth method (Fama and MacBeth, 1973) as follows. First, we estimate security betas using time series regressions with 3-year rolling windows. Second, we sort firms into 20 value-weighted portfolios based on their cyber beta. We compute the factor exposures of the portfolios and standardize the portfolio betas for economic interpretation. Finally, we estimate gammas using cross-sectional regressions of the portfolio returns on their lagged factor exposures.

Next, we use the time series approach of Gibbons et al. (1989) (hereafter, GRS) to test for portfolio efficiency. The GRS statistics allow testing whether the pricing errors are jointly equal to zero when using a model with several traded factors. We also use 20 portfolios built on factor betas as test assets. We implement the 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.

We 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 to price 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,

$$R_t = \alpha + \beta F_t + \epsilon_t, \;\; \epsilon_t \sim N(0, \Sigma), \tag{1}$$

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,

$$ML = ML_U(F|Mkt) \times ML_R(F^*|Mkt, F) \times ML_R(R|Mkt, F, F^*), \tag{2}$$

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,

$$\begin{aligned} ML_U &= |Y'Y|^{-N/2}|S|^{-(T-K)/2}Q \\ ML_R &= |Y'Y|^{-N/2}|S_R|^{-(T-K)/2}, \end{aligned} \tag{3}$$

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

$$Q = \left(1 + \frac{a}{a+k}(W/T)\right)^{-(T-K)/2} \left(1 + \frac{k}{a}\right)^{-N/2}, \tag{4}$$

where $a = {}^{(1+Sh(F)^2)}/_{T}$, $k = {}^{(Sh^2_{max} - Sh(Y)^2)}/_{N}$, W is the GRS F-statistic times ${}^{NT}/_{(T-N-K)}$,

$Sh(Y)^2$ 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, we 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,

$$P(M_j|D) = \frac{ML_j \times P(M_j)}{\sum_i ML_i \times P(M_i)}, \tag{5}$$

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 2, is the same for all combinations of F and F^*. Hence, it also cancels out in the division and can be omitted.

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

4. Results

4.1. Cyber risk measure

Table 3 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 our 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, we are confident that our measure is orthogonal to other characteristics known to price stock returns.

[Insert Table 3 here.]

4.1.1. Time series and industry properties

Figure 6 presents the cross-sectional average cyber risk for every year in the study sample. We observe a monotonic positive time trend in line with the results of Florackis et al. (2023) and Jamilov et al. (2021).

[Insert Figure 6 here.]

Figure 7 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 sectors such as "Oil and Gas" and "Chemicals", that traditionally rely less on technology systems have lower scores. Nonetheless, the variation of cyber risk across industries remains limited, similar to the overall cyber risk distribution.

[Insert Figure 7 here.]

4.1.2. Determinants of firm-level cyber risk

To investigate the dependence of cyber risk on firm characteristics, we perform two regressions, presented in Table 4. In Model 1, we control for year- and firm-fixed effects, and in Model 2, we control for year- and industry-fixed effects. In both models, firm age has a statistically significant negative coefficient at the 1% level, reinforcing the view that younger firms have higher cyber risk. The book-to-market coefficient is negative and significant at the 1% level, meaning 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.

[Insert Table 4 here.]

4.2. Univariate portfolio sorts

We sort firms into portfolios based on their cyber risk and study the returns on the portfolios. We use a firm characteristic approach based on the results of Daniel and Titman (1997), who argue that characteristics and not covariances determine expected returns. More precisely, we assign firms to five portfolios based on the cyber risk score of their most recent 10-K statement. We rebalance the portfolios quarterly to allow for listings and delistings and incorporate information from new 10-K statements. We build five value-weighted portfolios, where Portfolio 1 (5) is the low (high) cyber risk portfolio.

4.2.1. Full sample

We track the performance of the portfolios from January 2009 until December 2022. Figure 8 shows the evolution of the cumulative returns of the market and the five portfolios. We observe that the higher the cyber risk of the portfolio, the higher the cumulative returns. Portfolio 5 significantly outperforms the market.

[Insert Figure 8 here.]

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

[Insert Table 5 here.]

4.2.2. Before and after Florackis et al. (2023) was first released

We implement the same analysis as in Section 4.2.1 above, but stopping the study period at the time of 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 A4 presents the results using the period before the publication. We 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 A5 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 cyberattack in August 2021 during which more than 76.6 million current and former customers' information had been accessed[10]. 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[11]. 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 essential to note that the study sample after the publication is much smaller, and as a result, the observations could be spurious.

4.3. Double sorts

We also perform double sorts, where we first sort on a firm characteristic other than cyber risk and then sort the resulting portfolios again on cyber risk. Similarly to the previous section, we build value-weighted portfolios that we rebalance quarterly. We use three double sorts, first sorting on market beta, book-to-market ratio, or firm size and then sorting again on cyber risk. Table A6 shows the average excess returns of each portfolio. The factor structure of cyber risk is more prevalent among large firms, low to medium book-to-market firms, and medium beta firms. The long-short portfolio has positive excess returns for most of the portfolio sorts. This further supports the view that cyber risk captures a variation in average returns by controlling for market beta, firm size, and book-to-market value.

4.4. Fama-Macbeth regressions

Table 6 presents the results of Fama-Macbeth regressions. Model 1 only includes the market factor. The coefficient on the market is insignificant, and the average adjusted R-squared is small, showing that the CAPM can not price the cyber beta-sorted portfolios. In Model 2, we use the cyber risk proxy, and as shown, the risk premium is statistically significant, and the average adjusted R-squared increases from virtually 0 to 0.13. 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.

[Insert Table 6 here.]

[10]Available at:https://www.t-mobile.com/news/network/cyberattack-against-tmobile-and-our-customers
[11]Available at: https://cyberscoop.com/ransomware-treasury-cryptocurrency-sanctions/

Table A7 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 still significant at the 10% level.

The previous results are obtained using portfolios sorted on cyber risk. As highlighted by Lewellen, Nagel, and Shanken (2010), asset pricing tests, such as Fama-Macbeth regressions, can be improved by using additional portfolios sorted on other firm characteristics. Following their recommendations, we expand the set of test assets with portfolios sorted by industry. Table A8 presents the results. The cyber risk premium of 0.176% per month is still significant at the 5% level.

4.5. GRS test

We implement the GRS test as follows: we 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. We report the results and repeat this procedure, sorting portfolios on market beta, firm size, and book-to-market ratio.

Table 7 presents the results. 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, we 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. This is not the case when sorting on market beta. However, we can not reject the null hypothesis for either model.

[Insert Table 7 here.]

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

4.6. 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 Section 3.5 allows us to determine the combination of factors that are best in terms of pricing returns. Figure 9 presents the posterior probabilities of the five 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.

[Insert Figure 9 here.]

Figure 10 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 sample's end. Unlike the remaining factors, the investment and operating profitability factors also have very high cumulative probabilities.

[Insert Figure 10 here.]

Finally, we study the sensitivity of the model probabilities to the prior multiple. We repeat the analysis using three other values of prior multiple: 1.25, 2, and 3 (similarly to Barillas and Shanken (2018)). Table 8 reports the posterior model probabilities at the end of the sample for the top five models for each prior multiple. We 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.

[Insert Table 8 here.]

4.7. Robustness tests

4.7.1. Long-run cyber risk

By design, the cyber risk computed in Section 3.4.4 depends only on each company's most recent 10-K statement. It is 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 cybersecurity in year T and not having been attacked itself, the firm could decide not to talk about cyber risks in its following 10-K statements (in years $T + i$) or not as much as in year T, even though it still has similar cyber risks. The previously computed cyber risk measure would miss these cases as it has no memory.

Using the cyber risk defined in Section 3.4.4, we compute the expanding average cyber risk score to study the long-run cyber risk of firms. The long-run cyber risk score in year T is the average of its simple cyber risk scores from 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 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 or software), minimize its future cyber risk, and not discuss this risk in its subsequent 10-K statements. The low cyber risk scores in the upcoming years represent the firm's reality and should not be discarded.

We repeat the portfolio sorts from Section 4.2 using the long-run cyber risk to sort firms. Table A9 presents the results. We observe no significant change from Table 5, and the long-short portfolio remains significant at the 5% or 10% level. These results indicate that firms incorporate all available information in their newest 10-K statements regarding their cyber risk; hence, incorporating information from past statements does not improve the estimation of the cyber risk.

4.7.2. Controlling for cybersecurity firms

The cyber risk score constructed in Section 3.4.4 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[12].

As there is no dedicated cybersecurity industry classification, we identify cybersecurity firms using the HACK ETF[13]. As explained in the fund's description, this ETF invests in companies providing cybersecurity solutions, including hardware, software, and services. As cybersecurity providers, these firms are expected to discuss cybersecurity in their 10-K statements extensively, resulting in a false high cyber risk score. Indeed, these firms have an average score of 0.59, which is in the top 3% of cyber risk scores. We repeat the analysis from Section 4.2, and we exclude the holdings of the HACK ETF from the universe of firms.

Table A10 presents 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, we observe that the t-statistics on Portfolios 4 and 5 also increase. These results support the view that cyber risk is priced.

4.8. Comparison with the work of Florackis et al. (2023)

In order to compare our cyber risk measure to the one of Florackis et al. (2023), we start by computing the correlation between the two measures and obtain 0.34.[14] Encouragingly, the correlation is positive but relatively low, which shows that our measure is novel.

The methodology of this paper brings several improvements to the one used by Florackis et al. (2023). First, the Paragraph Vector model is more potent in capturing cyber risk-related content in 10-K statements than a dictionary approach. Indeed, the dictionary will never contain all the words needed to identify all cyber risk-related text. The dictionary used by Florackis et al. (2023) can identify a large number of these texts unless the authors

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

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

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

of the 10-K statements use slightly different wording than the others, which is not uncommon. For example, table 9 shows three paragraphs that were missed by their approach and consequently the firm-year observations got a cyber risk score of 0. Using the methodology presented in this paper, these 10-K statements were given an above-average cyber risk score.

[Insert Table 9 here.]

Furthermore, as mentioned in the literature review, the dictionary approach leaves many firm-year observations with cyber risk scores of zero. However, as shown by the examples in table 9, some of those firms are incorrectly identified as low-risk firms. To investigate this further, we use our cyber risk measure to sort firms with a zero cyber risk score (using the measure of Florackis et al. (2023)) into three portfolios. The average excess returns and alphas of the portfolios are shown in table 10 below. The average excess returns and alphas increase with the cyber risk (as measured using our indicator) and the long-short portfolio has economically significant excess returns and alphas.

[Insert Table 10 here.]

5. Conclusion

This paper implements a doc2vec model to estimate firms' cyber risk based on their 10-K statements. We then use this cyber risk measure in various asset pricing tests. Our results support the view that cyber risk is priced in the cross-section of stock returns. Indeed, a long-short strategy on cyber risk sorted portfolios has a positive and statistically significant alpha compared to traditional factor models and an average monthly excess return of 0.56%. We also perform this analysis limited to the sub-periods before and after the first release of Florackis et al. (2023). While the long-short strategy has statistically significant excess returns and alphas at the 1% level before the release, it has negative average excess returns after. We investigate the interactions of cyber risk with the other sources of risks by performing double sorts. The results show that cyber risk captures a variation in average returns by controlling for market beta, firm size, and book-to-market value. Furthermore, we show that cyber risk has a significant premium using Fama-Macbeth regressions. Using the GRS test of Gibbons et al. (1989) and the methodology from Barillas and Shanken (2018), we show that the cyber risk factor helps to price stocks and is present in the five most likely factor models. We also compute the long-run cyber risk as the expanding average of the cyber risk. We perform the portfolio sorts and observe that the results are very similar. Hence, we conclude that firms incorporate all available information about cyber risk in their newest 10-K statement. We exclude cybersecurity firms from the sample and perform the portfolio sorts.

We find that the average monthly returns of the two high cyber-risk portfolios increase, and the others are left unchanged, supporting the view that the alpha is due to the cyber risk. Finally, we compare our measure to the one of Florackis et al. (2023) and conclude that our method performs better at capturing cyber risk.

This research can be extended in two directions. First, we aim to repeat the study for the stock markets of different countries. Second, we intend 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 close to the risk under scrutiny.

References

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.

Daniel, K., Titman, S., 1997. Evidence on the characteristics of cross sectional variation in stock returns. Journal of Finance 52, 1–33.

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.

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., 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.

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`

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

Jiang, H., Khanna, N., Yang, Q., Zhou, J., 2023. The cyber risk premium. Management Science Forthcoming.

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 cyberattacks 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.

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

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96, 175–194.

Liu, J., Marsh, I. W., Xiao, Y., 2022. Cybercrime and the cross-section of equity returns. Avaiable at: http://dx.doi.org/10.2139/ssrn.4299599

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.

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.

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

Figures and Tables

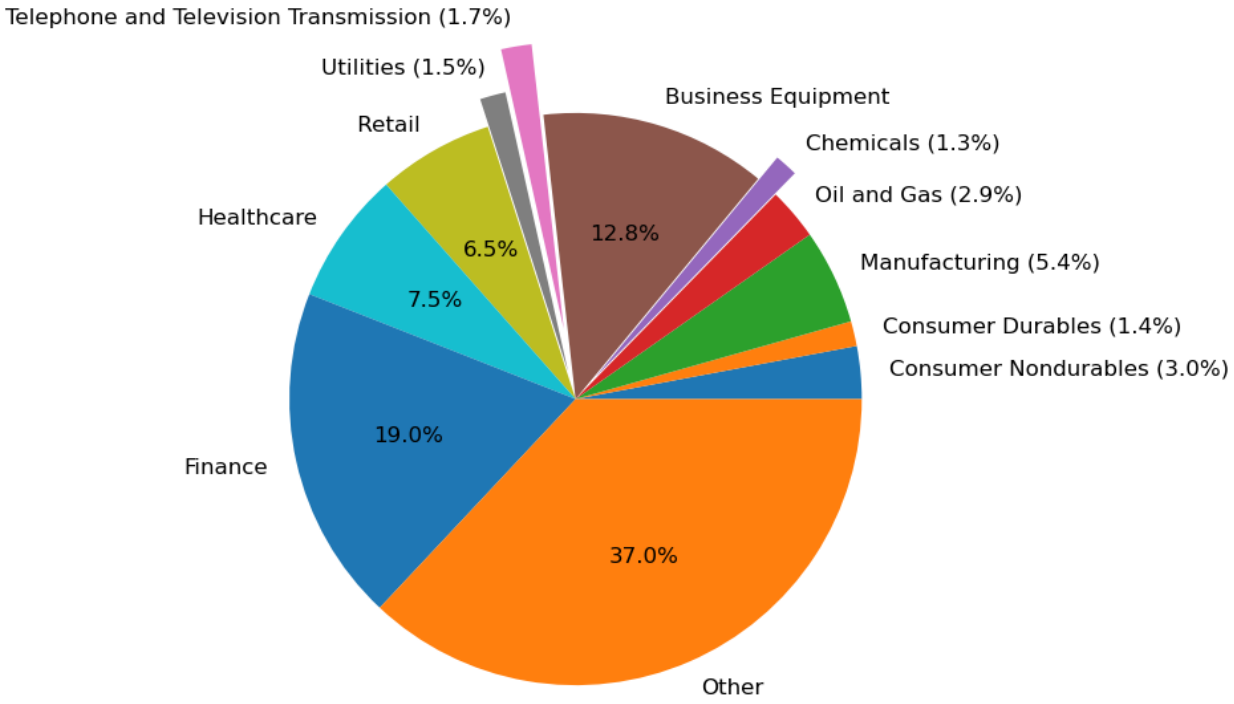

Fig. 1: **Industry distribution**

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.

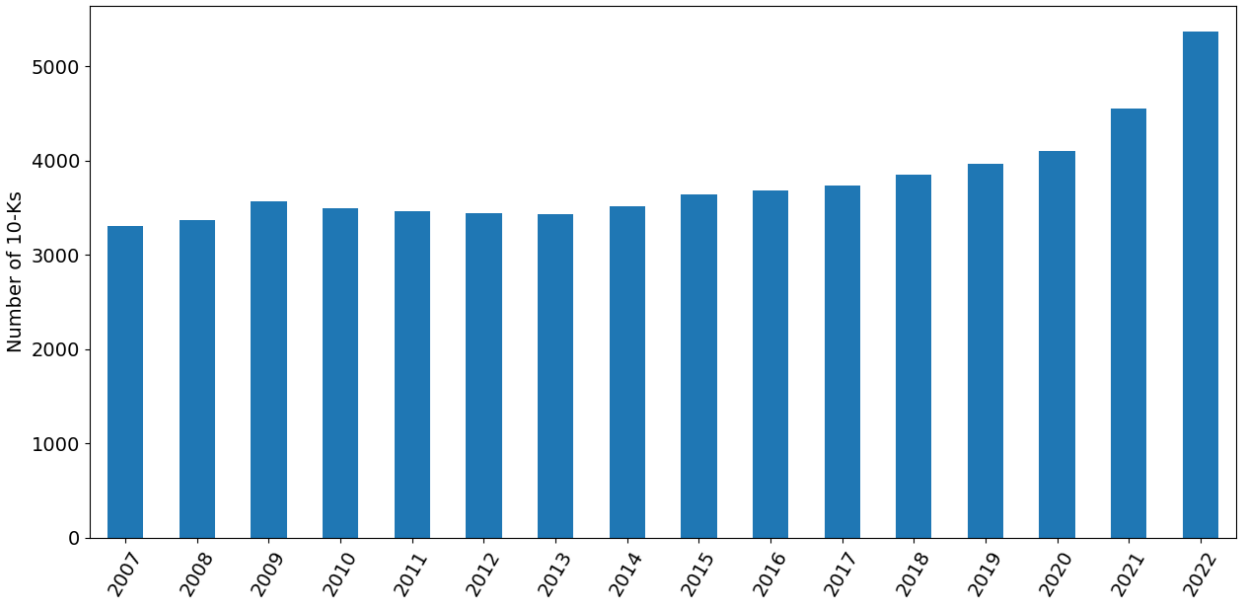

Fig. 2: **Number of 10-Ks per year**

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

Tactic

Tactics represent the "why" of an ATT&CK technique or sub-technique. It is the adversary's tactical goal: the reason for performing an action. For example, an adversary may want to achieve credential access, collect data or run malicious code.

Technique

Techniques represent "how" an adversary achieves a tactical goal by performing an action. For example, an adversary may dump credentials to achieve credential access or may use social engineering to run malicious code.

Sub-technique

Sub-techniques describe the different types of a technique. For example, when trying to run malicious code for social engineering, the adversary may use a malicious link, a malicious file or a malicious image for execution.

Fig. 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 1: **Examples of sub-technique descriptions from MITRE ATT&CK**

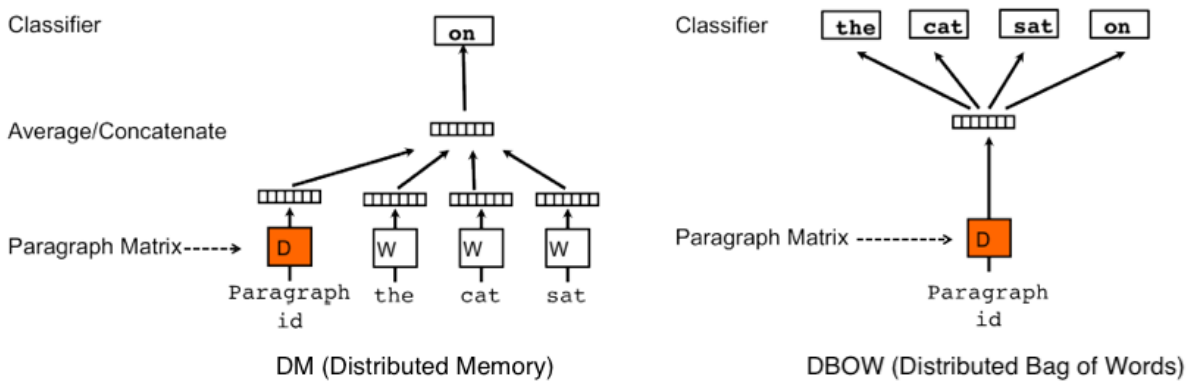

Fig. 4: **"Paragraph Vector" model versions**

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

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 2: **doc2vec parameters**

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

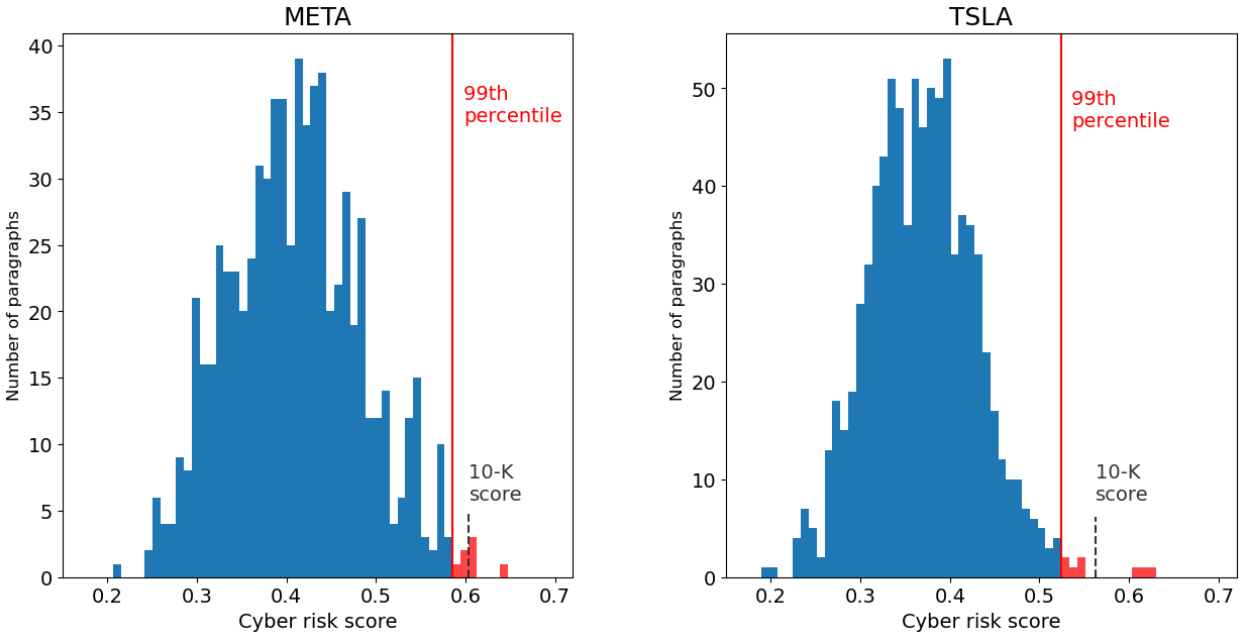

Fig. 5: **Paragraph level score distributions for Meta Platforms and Tesla**

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.

	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 3: **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 A3.

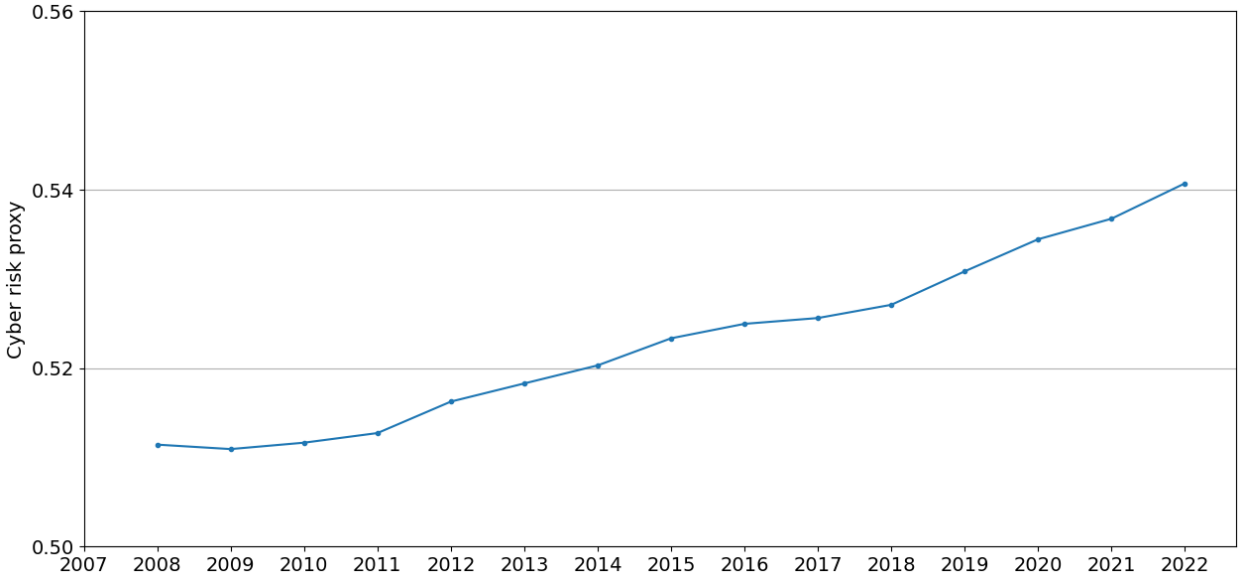

Fig. 6: **Evolution of the average cyber risk across all firms**

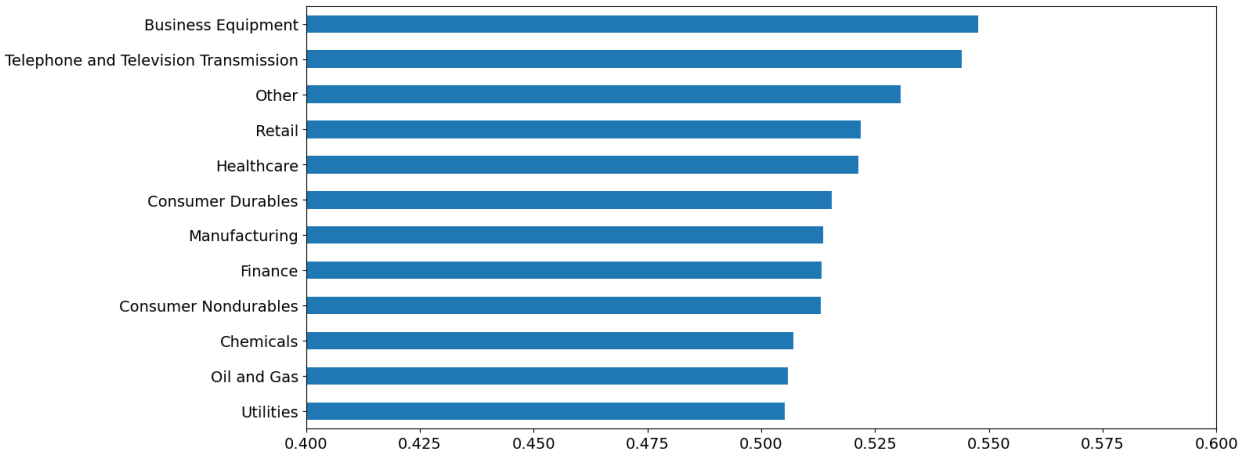

Fig. 7: **Average value of the cyber risk across industries**

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 4: **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 A3.

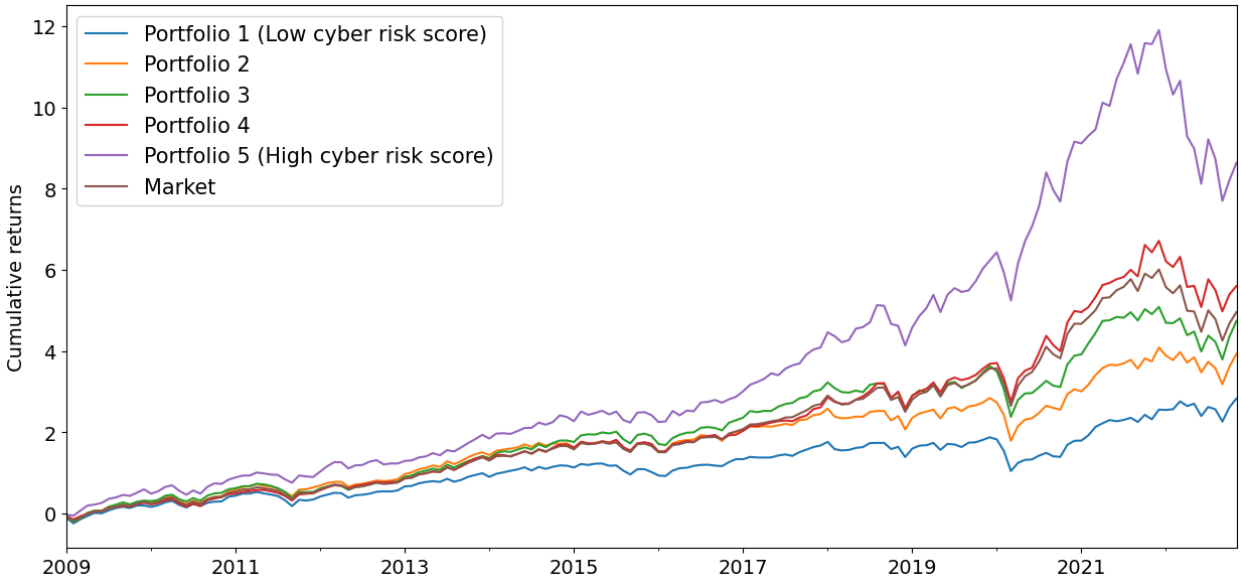

Fig. 8: **Cyber risk sorted portfolio cumulative returns**

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. One unit of the ordinate axis unit is 100% return.

	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	-
Sharpe Ratio	0.637	0.775	0.801	0.892	1.037	0.628
Treynor Ratio	0.031	0.037	0.038	0.042	0.050	3.283
Sortino Ratio	0.942	1.199	1.263	1.462	1.790	2.733

Table 5: **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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio 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

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 6: **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 t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

	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 7: **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 Section 4.2.1.

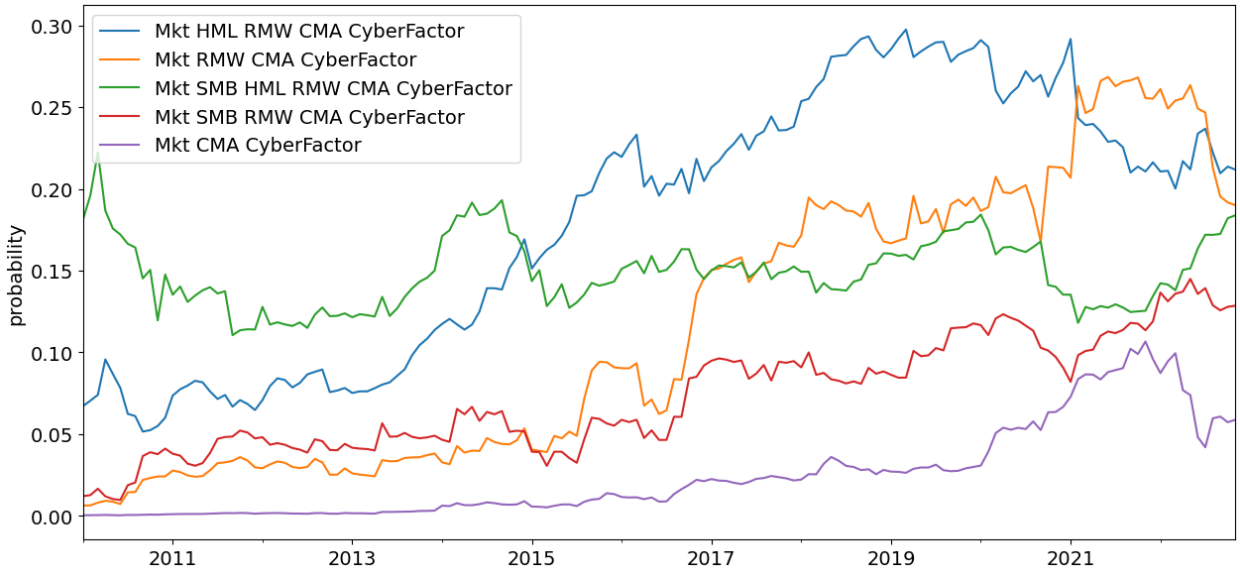

Fig. 9: **Factor model posterior probabilities**

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 Section 4.2.1. Prior Multiple = 1.5

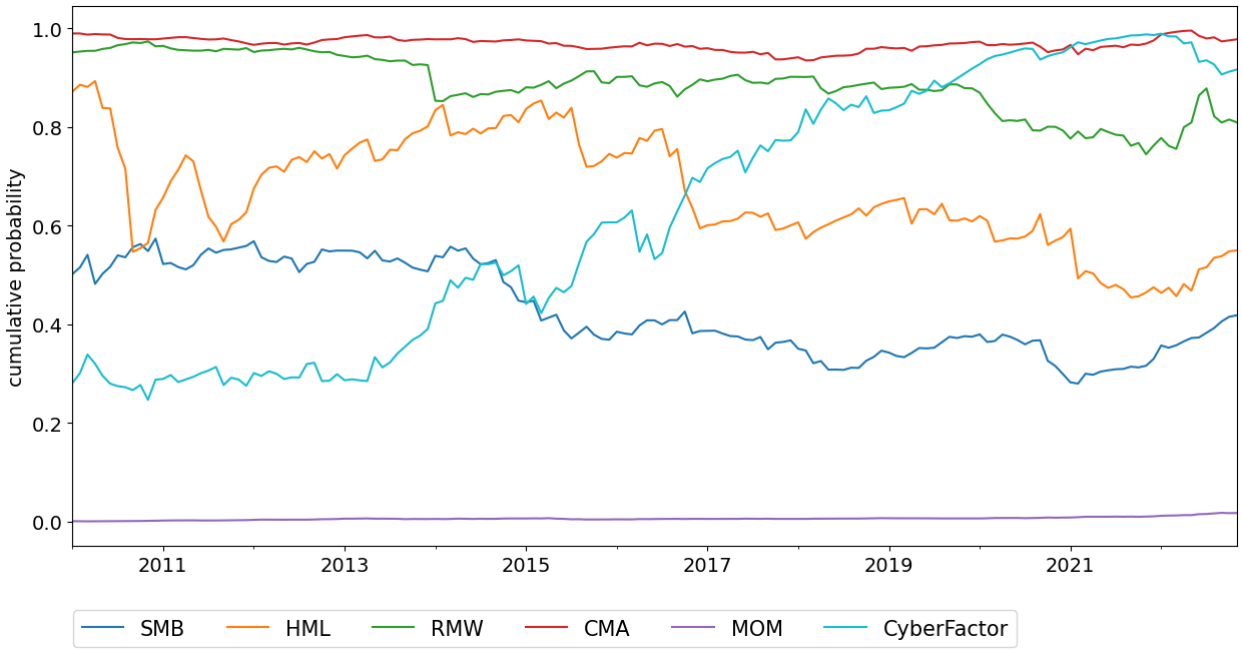

Fig. 10: **Cumulative posterior factor probabilities**

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 Section 4.2.1. Prior Multiple = 1.5

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 CyberFactor	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 8: **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 Section 4.2.1

Ticker	Filling year	Paragraph
STX	2008	System failures caused by events beyond our control could adversely affect computer equipment and electronic data on which our operations depend. Our operations are dependent upon our ability to protect our computer equipment and the electronic data stored in our databases from damage by, among other things, earthquake, fire, natural disaster, power loss, telecommunications failures, unauthorized intrusion and other catastrophic events. As our operations become more automated and increasingly interdependent, our exposure to the risks posed by these types of events will increase. While we continue to improve our disaster recovery processes, system failures and other interruptions in our operations could have a material adverse effect on our business, results of operations and financial condition.
MANH	2014	Our software may contain undetected errors or "bugs" resulting in harm to our reputation which could adversely impact our business, results of operations, cash flow, and financial condition. Software products as complex as those offered by us might contain undetected errors or failures when first introduced or when new versions are released,. Despite testing, we cannot ensure that errors will not be found in new products or product enhancements after commercial release,. Any errors could cause substantial harm to our reputation, result in additional unplanned expenses to remedy any defects, delay the introduction of new products, result in the loss of existing or potential customers, or cause a loss in revenue. Further, such errors could subject us to claims from our customers for significant damages, and we cannot assure you that courts would enforce the provisions in our customer agreements that limit our liability for damages. In turn, our business, results of operations, cash flow, and financial condition could be materially adversely affected.
MBCN	2017	Material breaches in security of bank systems may have a significant effect on the Company business. We collect, process and store sensitive consumer data by utilizing computer systems and telecommunications networks operated by both banks and third party service providers. We have security, backup and recovery systems in place, as well as a business continuity plan to ensure systems will not be inoperable. We also have security to prevent unauthorized access to the system. In addition, we require third party service providers to maintain similar controls. However, we cannot be certain that these measures will be successful. A security breach in the system and loss of confidential information could result in losing customers' confidence and thus the loss of their business as well as additional significant costs for privacy monitoring activities.

Table 9: **Examples of paragraphs missed by the dictionary approach of Florackis et al. (2023)**

	Value Weighted Portfolios			
	L		H	H-L
	P1	P2	P3	P3-P1
A. Portfolios sorted by cyber risk				
Average excess return	**0.71**	**0.88****	**1.24*****	**0.53****
	[1.83]	[2.25]	[3.37]	[2.00]
CAPM alpha	**-0.58*****	**-0.35**	**-0.07**	**0.51***
	[-2.75]	[-1.11]	[-0.35]	[1.73]
FFC alpha	**-0.40***	**-0.23**	**-0.05**	**0.35**
	[-1.89]	[-0.82]	[-0.37]	[1.38]
FF5 alpha	**-0.37***	**-0.27**	**0.01**	**0.38**
	[-1.73]	[-0.97]	[0.09]	[1.54]
B. Characteristics				
Number of firms	248.5	248.9	248.3	-
Cyber risk	0.486	0.504	0.534	-
Sharpe Ratio	0.489	0.645	0.856	0.567
Treynor Ratio	0.023	0.030	0.039	3.976
Sortino Ratio	0.735	1.071	1.463	2.349

Table 10: **Average monthly excess returns and alphas (in percent), zero-risk firms using the measure of Florackis et al. (2023)**

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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio 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 2018

	Value Weighted Portfolios					
	L P1	P2	P3	P4	H P5	H-L P5-P1
A. Portfolios sorted by cyber risk						
Average excess return	**0.97*****	**1.05*****	**1.15*****	**1.19*****	**1.48*****	**0.51**
	[3.61]	[4.01]	[3.99]	[4.10]	[4.24]	[1.62]
CAPM alpha	**-0.15**	**0.00**	**-0.01**	**0.07**	**0.33**	**0.48**
	[-0.81]	[-0.02]	[-0.04]	[0.85]	[1.54]	[1.26]
FFC alpha	**-0.09**	**0.05**	**0.04**	**0.08**	**0.24***	**0.33***
	[-0.98]	[0.46]	[0.45]	[1.16]	[1.82]	[1.69]
FF5 alpha	**-0.10**	**-0.03**	**0.05**	**0.05**	**0.25***	**0.35****
	[-1.40]	[-0.38]	[0.51]	[85]	[1.90]	[2.10]
B. Characteristics						
Number of firms	592.9	592.3	592.1	592.1	592.6	-
Cyber risk	0.482	0.496	0.507	0.521	0.565	-
Sharpe Ratio	0.702	0.820	0.829	0.883	1.029	0.571
Treynor Ratio	0.034	0.039	0.039	0.042	0.051	2.027
Sortino Ratio	1.058	1.305	1.346	1.417	1.777	2.818

Table 11: **Average monthly excess returns and alphas (in percent), omitting Item 1A**

During construction of the cyber risk measure, Item 1A is omitted from the 10-K statements. 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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio 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

Appendix

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

Table A1: **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 A2: **Top scoring paragraphs from the doc2vec validation sample**

The paragraphs are shown after preprocessing (as described in section 3.4.1). 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 A2: **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 A3: **Variable definitions**

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

	Value Weighted Portfolios					
	L P1	P2	P3	P4	H P5	H-L 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	-
Sharpe Ratio	0.511	0.752	0.807	0.965	1.272	1.157
Treynor Ratio	0.024	0.034	0.037	0.043	0.060	-0.714^{+}
Sortino Ratio	0.723	1.137	1.257	1.571	2.393	4.319

Table A4: **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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio of the portfolios. Newey-West 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).

$^{+}$*Note that the negative Treynor ratio is due to the negative correlation between the long-short and the market portfolio in the sample, and is hence not a sign of inferior risk-return characteristics.*

	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	-
Sharpe Ratio	1.374	1.009	0.758	0.873	0.357	-1.287
Treynor Ratio	0.088	0.061	0.045	0.051	0.021	0.092
Sortino Ratio	2.798	1.830	1.280	1.665	0.529	0.640

Table A5: **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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio of the portfolios. Newey-West 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)

	Value Weighted Portfolios				
	Cyber Q1 (Low)	Cyber Q2	Cyber Q3	Cyber Q4	Cyber Q5 (High)
A. sorted on market beta					
Beta Q1 (Low)	0.94	1.01	0.85	1.09	0.91
Beta Q2	1.04	1.03	1.23	1.25	1.55
Beta Q3	0.98	1.22	1.25	1.36	1.39
Beta Q4	1.36	1.19	1.38	1.29	1.96
Beta Q5 (High)	1.61	1.71	1.79	1.86	1.72
B. sorted on book-to-market					
BM Q1 (Low)	0.90	1.09	1.22	1.28	1.54
BM Q2	1.02	1.09	1.32	1.23	1.36
BM Q3	0.99	1.04	1.23	1.18	1.51
BM Q4	1.04	1.17	1.29	1.25	1.37
BM Q5 (High)	1.87	1.77	1.73	1.64	1.71
C. sorted on size					
Size Q1 (Small)	1.33	1.25	1.45	1.77	1.24
Size Q2	1.35	1.54	1.23	1.18	1.38
Size Q3	1.43	1.39	1.19	1.33	1.47
Size Q4	1.16	1.21	1.29	1.24	1.30
Size Q5 (Large)	0.86	1.00	1.14	1.25	1.52

Table A6: **Average monthly excess returns (in percent)**

Portfolios are first formed on market beta, book-to-market or size and then on cyber risk. Period: January 2009–December 2022

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 A7: **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 Fama-French 12 industry classification. 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 t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

Dependent variable: Monthly Portfolio returns	
	(7)
Market	**-0.033**
	[-0.578]
Cyber risk	**0.176****
	[2.103]
HML	**-0.033**
	[-0.293]
SMB	**-0.042**
	[-0.385]
RMW	**-0.086**
	[-1.150]
CMA	**0.011**
	[0.076]
Constant	**1.362*****
	[5.015]
$\overline{R2_{adj}}$	0.295
MAPE	1.243

Table A8: **Fama-MacBeth regressions with additional test assets**

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 set of test assets is expanded with portfolios sorted along industries. The industries correspond to the Fama-French 12 industry classification. $\overline{R2_{adj}}$ is the average adjusted R-squared and MAPE is the mean average pricing error. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

	Value Weighted Portfolios					
	L P1	P2	P3	P4	H P5	H-L 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**	**0.01**	**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	-
Sharpe Ratio	0.610	0.818	0.816	0.918	1.027	0.638
Treynor Ratio	0.030	0.039	0.038	0.044	0.050	2.572
Sortino Ratio	0.893	1.305	1.294	1.525	1.768	2.643

Table A9: **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, the average long-run cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio of the portfolios. Newey-West 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 P1	P2	P3	P4	H P5	H-L 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	-
Sharpe Ratio	0.637	0.772	0.796	0.903	1.049	0.641
Treynor Ratio	0.031	0.037	0.038	0.042	0.051	4.408
Sortino Ratio	0.942	1.196	1.253	1.483	1.812	2.759

Table A10: **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, the average cyber risk, the annualized Sharpe ratio, the annualized Treynor ratio, and the annualized Sortino ratio of the portfolios. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Period: January 2009-December 2022

Score	Paragraph	Ticker	Sub-technique (Tactic)
-0.097	The increase in income from operations during 2021 was primarily the result of higher net sales and production volumes and improved margins, which benefited from positive pricing impacts and a favorable sales mix that offset substantial inflationary material and freight cost pressures along with higher manufacturing costs. Net income per diluted share was favorably impacted by the reversal of valuation allowances previously established against our deferred tax assets in both the United States and Brazil during 2021	AGCO	SSH Authorized Keys (Persistence)
-0.092	In addition, we generally would expect to be able to recover a significant portion of the amounts paid under such guarantees from the sale of the underlying financed farm equipment, as the fair value of such equipment is expected to offset a substantial portion of the amounts paid. We also guarantee indebtedness owed to certain of our finance joint ventures if dealers or end users default on loans.	AGCO	Fast Flux DNS (Command and Control)
-0.099	In December 2021, the Company completed the sale of its Le Parfait brand in Europe and a previously closed plant in the Americas. Gross proceeds on these divestitures were approximately $113 million and the related pretax gains (including costs directly attributable to the sale) were approximately $84 million ($70 million after tax) in 2021. The pretax gains were recorded to Other income (expense), net on the Consolidated Results of Operations. In January 2021, the Company completed the sale of its plant in Argentina.	OI	GUI Input Capture (Collection)
-0.088	In 2020, the Company recognized a net gain (including costs directly attributable to the sale of ANZ and subject to post-closing adjustments) on the divestiture of approximately $275 million, which was reported on the Other income (expense), net line in the Consolidated Results of Operations. In addition, at closing, certain subsidiaries of the Company entered into certain ancillary agreements with Visy and the ANZ businesses in respect of the provision of certain transitional and technical services to the ANZ businesses.	OI	Scheduled Transfer (Exfiltration)

Table A11: **Examples of negative scoring paragraphs**

Examples of paragraphs with negative similarities, as measured by cosine similarity. Sub-technique refers to the MITRE sub-technique the paragraph is most dissimilar to. The displayed score is the cosine similarity between the paragraph and the previously mentioned sub-technique. The two companies, AGCO Corporation (AGCO) and O-I Glass Inc (OI), were chosen because of their low cyber risk scores. The 10-Ks are the ones filed in 2022.