Cyber risk and the cross-section of stock returns

Daniel Celeny* and Loïc Maréchal†

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 score 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, Cyber-Defence Campus, armasuisse S+T.loic.marechal@unil.ch

This document results from a research project funded by the Cyber-Defence Campus, armasuisse Science and Technology. Its initial version won the first prize in the 2024 edition of the Natixis Prize for Best Master Thesis in Quantitative Finance. Preprints of early-stage versions of this research are available on SSRN and arXiv. We appreciate helpful comments from seminar participants at the Cyber Alp Retreat 2023 and the Workshop on the Economics of Information Security 2024. We also thank Pierre Collin-Dufresne and Michel Dubois for their invaluable comments. Our code will be available at `https://github.com/loicym/Cyber-risk_and_the_cross-section_of_stock_returns` Corresponding author: Loïc Maréchal e-mail: loic.marechal@unil.ch, University of Lausanne, Internef building, Chamberonne district, 1015 Lausanne, Switzerland

1. Introduction

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 cyber risk score shows a monotonic increasing trend, moving from 0.51 to 0.54. In contrast, the cross-sectional distribution of that score is exceptionally narrow (standard deviation of 0.03). We compare our cyber risk score across Fama-French 12 industries and find results aligned with the view that "Business Equipment" and "Telephone and Television Transmission" are the riskiest and "Oil and Gas" and "Utilities", the safest.

The cyber risk sorted long-short portfolio, which invests in high cyber risk stocks and shorts low cyber risk stocks, has an average monthly excess return of 0.56% (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 performed particularly well before the first mention of a cyber risk premium on SSRN in November 2020 by Florackis, Louca, Michaely, and Weber (2023), with an average annual excess return of 11.88%. It is statistically significant at the 1% level. Double sorts confirm that cyber risk captures a variation in stock returns when controlling for market beta, size, book-to-market, momentum, investment, and operating profitability.

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 estimate, 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 score 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 cyber score to the one obtained by Florackis et al. (2023) with the dictionary approach used. While the two estimates are only weakly positively correlated at 34%, we argue that our score performs better. First, it allows for capturing variations in cyber risk exposure of firms previously assigned zero cyber risk when the dictionary approach was used. Second, our score enables us to treat any firm-related text and not only dedicated risk sections such as Items 1. A for 10-Ks. We also argue that previous methods to extract cyber risk scores from financial documents must be revised since the dictionary approach may miss the context and not allow for capturing high-level semantics. Moreover, these methods require training based on the actual realization of ex-ante (Florackis et al., 2023) or ex-post (Jiang, Khanna, Yang, and Zhou, 2023) cyberattacks. In the former case, this raises doubts about whether the effect captured is a "post-attack recovery"; in the latter, there is no possible real-life implementation achievable, and even with great care, one could not discard forward-looking bias. In contrast, we are confident that our method is immune from these biases and would only capture the latent cyber risk perceived by investors. We demonstrate it conceptually and with results from more powerful asset pricing methods than those previously used to test the performance of the cyber risk score. Altogether, we find strong support for cyber risk being priced in the cross-section of stock returns.

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

2. Literature review

2.1. The costs of cyber risk

Amir, Levi, and Livne (2018) compare firms with voluntary disclosures and those with-

holding information. They find that while the formers are associated with a 0.7% decline in share value, the latter experience a 3.6% decline, the month during which the attack is discovered. Gordon, Loeb, and Sohail (2010) assess the firm value of voluntary cybersecurity disclosures. They find positive correlations between voluntary disclosure and both firm value and liquidity.

Anderson, Barton, Böhme, Clayton, Eeten *van*, Levi, Moore, and Savage (2013) perform a systematic analysis of the costs of cybercrimes. The typical yearly cost for a citizen lies between hundreds of USD (for tax and welfare fraud) and a few cents (for "new crime" such as botnets). Revisited version by Anderson, Barton, Böhme, Clayton, Ganan, Grasso, Levi, Moore, and Vasek (2019) shows that payment frauds have doubled over the seven years separating initial studies, along with a significant fall in average costs for citizens. Bouveret (2018) estimates an average cyberattack loss of USD 97 billion at the country level and a value-at-risk in the USD 147-201 billion range. Closely, Romanosky (2016) finds that the cost distribution is heavily skewed, with an average cost of USD 6 million and a median cost of USD 170,000.

Much literature on the economic impact of cyber incidents involves event studies. For instance, Gordon, Loeb, and Zhou (2011) find that news about information security breaches significantly negatively affects stock returns, although they also document a downward trend of this effect over time. Similarly, Campbell, Gordon, Loeb, and Zhou (2003) uncover a systematic adverse market reaction around cyber incidents but only when it involves confidential information. Similarly, studying data breaches Johnson, Kang, and Lawson (2017) find that, on average, stocks decrease by 0.37% around them, whereas Lending, Minnick, and Schorno (2018) uncover a long-term effect and identify one-year buy-and-hold abnormal returns of -3.5%. Studying the channels by which these abnormal returns around cyber incidents materialize, Tosun (2021); Kamiya, Kang, Kim, Milidonis, and Stulz (2021) attribute their effects to arise from shocks to firms' reputations. Finally, Andreadis, Kalotychou, Louca, Lundblad, and Makridis (2023) focus on the US municipal bonds market and estimates that a 1% increase in the number of cyberattacks leads to an increase in yields from 3.7 to 5.9 basis points. Conversely, Jensen and Paine (2023) find no immediate effect on bond yields of hacked towns. Furthermore, in the subsequent 24 months, they document the municipal bond yields to decline along with an increase in IT spending.

2.2. Information extraction from natural language

Natural language processing in the financial literature often uses sentiment, *i.e.* a concept that parses the single dimension of positive *vs.* negative content. To alleviate this concern, Loughran and McDonald (2011), while showing that "standard" dictionaries can not capture the sentiment in financial documents, develop negative and positive dictionaries and four others to extract other tones.[1] Similarly, Jegadeesh and Wu (2013) use positive and negative dictionaries to extract sentiment from 10-K statements that directly map to market reactions. Garcia (2013) uses ratios of positive and negative words from the New York Times news over a century to predict stock returns in the US, while Calomiris and Mamaysky (2019) employ a similar method to predict risk and returns in stock markets worldwide.

Dictionary-based methods may also capture more meaning than positive or negative sentiment. For instance, Bodnaruk, Loughran, and McDonald (2015) study 10-K statements with a dedicated lexicon identifying constraining words and build a measure uncorrelated with traditional financial constraint measures, yet accurately predicting liquidity events. Antweiler and Frank (2004) use a measure of stock disagreement on internet forums and identify that disagreement is positively associated with trading volume. Hassan, Hollander, *van* Lent, and Tahoun (2019) use word and bigram counts in quarterly earnings conference calls to capture the political risk of US-listed firms, whereas Sautner, *van* Lent, Vilkov, and Zhang (2023) use a similar approach in the same documents to capture climate change risk contexts. Finally, Hoberg and Phillips (2010) uses a measure of product market language proximity between firms in 10-K statements and finds that it predicts the likelihood of a future merger or acquisition between them.

Our study borrows from recent advances in the machine-learning treatment of texts. Le and Mikolov (2014) develop 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

[1] The other tones are uncertainty, litigious, strong modal, and weak modal.

applications. Mendsaikhan, Hasegawa, Yamaguchi, and Shimada (2019) use the Paragraph Vector algorithm with cybersecurity-specific training data to identify cybersecurity-related texts. Their training data combines online forums, news outlets, and vulnerability databases. Their model identifies cybersecurity texts with 83% accuracy. Finally, close to our research, Adosoglou, Lombardo, and Pardalos (2021) construct portfolios based on the semantic differences between two consecutive 10-K statements of each firm, thereby revisiting Cohen, Malloy, and Nguyen (2020). They find that cosine similarity is the most effective metric for neural network embeddings, such as the ones obtained using the Paragraph Vector algorithm.

2.3. Is cyber risk a priced factor?

Florackis et al. (2023) build a text-based cyber risk score using the risk-disclosure section of 10-K statements ("Item 1. A Risk Factors"). They restrict their analysis to cyber-risk-related sentences from this section, extracted with a match on a list of cyber-related keywords. They consider recently (-1 or -2 years before filing) hacked firms as a training sample and compute the cyber risk exposure of firms. This exposure is computed 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, Rey, and Tahoun (2023) perform a similar analysis using quarterly earnings calls worldwide from 2002 to 2021. Borrowing from the NLP approach of Hassan et al. (2019) to extract political risk, they construct a cyber risk score using the frequency of cyber-related keywords in earnings calls and find that cyber risk exposure is priced in stocks and options markets and predicts cyberattacks. Jiang et al. (2023) use a cyber dictionary from NIST to count the number of cyber-related words in Item 1. A. Combining it with firm characteristics, they perform a logit ridge regression where the dependent variable is the probability for a firm to experience a cyberattack in the future. We argue that their method has several drawbacks. First, because their calibration uses forward-looking data, their results could not be obtained in real life since an asset manager would have to wait until the realization of cyberattacks to calibrate their model.[2] Second, despite claiming no forward-looking bias, which is the case

[2]The chronology and synchronicity of predictors and cyberattacks is not very well defined in their re-

from the predictors' availability, they nonetheless use forward-looking information (cyberattack realizations). In an intertemporal CAPM framework, Liu, Marsh, and Xiao (2022) estimate firms' betas of stocks with rolling regressions *vs.* a cybercrime news attention proxy (Thomson-Reuters MarketPsych, TRMI). They find that high cybercrime-beta stocks generate 0.6% lower return than low cybercrime-beta stocks, which apparently contradicts the other studies. However, they capture sensitivities to contemporaneous realized cyberattacks, not latent cyber risk. The estimated beta measures firms' sensitivities to cyberattacks, not their probability of experiencing one.

We are only aware of these four studies focusing on cyber risk and its implications for asset pricing. While they all use a dictionary approach as a starting point, directly (Florackis et al., 2023; Jamilov et al., 2023), indirectly (Liu et al., 2022), or partially (Jiang et al., 2023), we argue that this method is unsuitable for estimating firm-level cyber risk exposure for several reasons. First, it leaves many firm-year observations with a cyber risk equal to 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. (2023)). Second, missing observations or measurement errors may imply building a common cyber factor from which firm-level betas are estimated instead of directly estimating it at the firm-level (see Liu et al., 2022).

3. Data and methodology

3.1. Market data

We download public equity data from Wharton Research Data Services (WRDS), using 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 A1. 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. We begin our analysis in 2007 to align with Florackis et al. (2023). Moreover, as Florackis et al. (2023) justify, it is unlikely that

search. However, we can only assume that from their statements "All the variables are measured prior to the year when the incidents happen".

the perceived (or actual) cyber risk was significant before this date. We also map firm-level information with the Fama-French 12 industry classification.[3]

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[4].

3.2. 10-K statements

To download 10-K statements, we use the index files from the SEC's Edgar archives[5]. 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 write 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 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 over the sample.[6] To avoid any forward-looking bias, we construct and rebalance portfolios quarterly based on all newly issued 10-Ks filed and available on the SEC Edgar archives before the rebalancing date.

3.3. Cybersecurity tactics

We use the MITRE ATT&CK[7] cybersecurity knowledge base as a reference for cybersecurity descriptions. This knowledge base was created in 2013 to document cybersecurity tactics, techniques, and procedures used by adversaries against particular platforms, such

[3] Appendix Figure A1 depicts the industry distribution.

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

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

[6] Appendix Figure A2 shows the number of 10-Ks filed per year. This number increased from 3,301 in 2007 to 5,370 in 2022.

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

as Windows or Google Workspace.[8] There are 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.

3.4. Methodology

3.4.1. Text preprocessing

We download 10-K statements from the SEC Archives as HTML files. We use the BeautifulSoup[9] 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 write a Python script that uses the wordfreq[10] and NLTK[11] 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 frequently appear in the texts, removing them allows us to focus on essential cybersecurity-related words .[12]

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 algorithm, proposed by Le and Mikolov (2014), is an extension of the word2vec algorithm (Mikolov, Chen, Corrado, and Dean (2013)). It aims to learn fixed-

[8] Appendix Figure A3 illustrates the structure of the knowledge base (Panel A) and provides a short description (Panel B).

[9] Available at: https://www.crummy.com/software/BeautifulSoup/

[10] Available at: https://pypi.org/project/wordfreq/

[11] Available at https://www.nltk.org/

[12] Stop-words contain no valuable information that could hint at the paragraph's content. For this reason, it is standard practice to remove them before using such models (see, *e.g.*, Mendsaikhan et al., 2019). We also attempted to train the models to keep stop-words, but they were then less powerful at identifying cyber-related paragraphs.

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

The algorithm 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 1 illustrates the two models.

[Insert Figure 1 here.]

Both models are unsupervised, as the paragraph vectors are learned from unlabelled data. We use the implementation by Gensim called doc2vec[13]. 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 specifications using this sample with various vector dimensions, epochs, and window sizes. The baseline for the hyperparameters is taken from Lau and Baldwin (2016) (see Table 1).

To choose the best specification, we compute the vector representations of the paragraphs from 10 randomly chosen 10-K statements from 2008 (validation sample) using each specification and compare the highest-scoring paragraphs between the models (the scoring algorithm is explained below). We choose the best specification where the proportion of the highest-scoring cyber risk-related paragraphs is the highest. Table 1 presents the parameters of the best-performing doc2vec specification, which is used for the remainder of this study. Appendix Table A2 gives an excerpt of top-scoring paragraphs from the validation sample (after pre-processing).

[Insert Table 1 here.]

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

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 estimate 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. The similarity between two paragraphs, with vector representations v_1 and v_2, is therefore computed as $sim = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$.

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 best-performing specification. 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, 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 (average of 638 paragraphs per 10-K statement). 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.

Since the cosine similarity takes values between -1 and 1, we only consider positive values for several reasons. First, interpreting a paragraph whose context is "opposite" to cybersecurity is not intuitive. Upon inspection of 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. (2023); Florackis et al. (2023).[14]

Figure 2 shows the distribution of the cyber risk scores of the paragraphs from the 10-K statements of Meta Platforms, Inc. and Tesla, Inc. filed in 2022. The paragraphs in red are the top 1% of paragraphs with the highest cyber risk scores. We compute the 10-K cyber risk scores as the average score of this highest percentile.[15] This yields a score of 0.605 for META and 0.563 for TSLA.

[Insert Figure 2 here.]

3.5. Asset pricing tests

We use two-step Fama-MacBeth regressions (Fama and MacBeth, 1973) in a multifactor asset pricing theory context (Ross (1976)) as follows. First, we estimate security betas on several factors using time series regressions with 2-year rolling windows. Second, we sort firms into 20 value-weighted portfolios based on their cyber risk. 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 and their value-weighted cyber risk score.

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 cyber risk scores and factor betas, in turn, 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

[14]We report in Appendix Table A4 an excerpt of these negative paragraphs to illustrate that they are orthogonal to the concept of cyber risk and not indicators of cybersecurity.

[15]We try, in turn, other scores, such as the firm average of the cyber risk score distribution, with virtually no changes. We retain the top percentile approach since, intuitively, these paragraphs are the most likely to discuss actual cybersecurity issues.

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 still need to be identified. The approach compares the relative success of the models in predicting the observed data. The method is based on Barillas and Shanken (2017) and 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 $a = {}^{(1+Sh(Y)^2)}/_{T}$, $k = {}^{(Sh^2_{max} - Sh(Y)^2)}/_{N}$, 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 score

Table 2 presents descriptive statistics of the cyber risk score 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. 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 score and firms' characteristics are small, except for Tobin's Q (0.23), risk section length (0.21), and 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 score is orthogonal to other

characteristics known to price stock returns.

[Insert Table 2 here.]

4.1.1. Time series and industry properties

Figure 3 presents the cross-sectional average cyber risk score 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. (2023).[16]

[Insert Figure 3 here.]

Figure 4 shows the average cyber risk by industry, using the Fama-French 12 industry classification (also see, Fama and French, 1997). Industries that rely on technology systems such as "Business Equipment" and "Telephone and Television Transmission" have high cyber risk. In contrast, sectors such as "Oil and Gas" and "Chemicals", which traditionally rely less on technology systems, have lower scores. Nonetheless, the dispersion of cyber risk across industries remains limited, similar to the overall one, and appears inversely proportional to the cyber risk magnitude.

[Insert Figure 4 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 3. In Model 1, we control for year- and firm-fixed effects; 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 are exposed to a 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 coefficient of intangible assets to total assets is positive and statistically significant at the 1% level. This supports the view that firms with more intangible assets, such as patents or software, have a higher cyber risk. Last, the risk section length coefficient

[16] We are aware that this increasing trend can either reflect the increasing actual cyber risk or the fact that 10-Ks authors are increasingly reporting it. Nonetheless, we argue that in both cases, this would translate into investors' increased perception of this risk, which would turn to the same outcome in terms of pricing.

is positive and significant at the 1% level in both models. We highlight, however, that the t-statistics are one order of magnitude lower than those found by Florackis et al. (2023) with t-statistics above 40 and 20 in their specifications. This, along with the readability coefficient flipping sign across models and an R-squared within which is lower than theirs for both models, support the view that we are less likely to capture other firms' characteristics or 10-K features and that our measure is novel and is better at capturing the latent cyber risk exposure of firms.

[Insert Table 3 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. We track the performance of the portfolios from January 2009 until December 2022. Figure 5 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. Specifically, Portfolio 5 significantly outperforms the market.

[Insert Figure 5 here.]

Table 4 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 remain statistically significant at the 5% level after controlling for the Fama and French (2015) five-factor model. In addition, we also observe that the Sharpe, Treynor, and Sortino ratios all experience monotonical growth as the cyber risk increases, reinforcing our views that the performance is not driven

by compensation for total volatility, market exposure, or downside risk, respectively.[17]

[Insert Table 4 here.]

4.3. Double sorts

We also perform double sorts, where we first sort five portfolios on an alternative firm characteristic and then sort the resulting portfolios on cyber risk. Akin to the previous section, we build value-weighted portfolios that we rebalance quarterly. We use three characteristics for the first sort: market beta, book-to-market ratio, and firm size. Table 5 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 specific variation in average returns by controlling for market beta, firm size, and book-to-market value.

[Insert Table 5 here.]

4.4. Asset pricing tests

4.4.1. Fama-Macbeth regressions

Table 6 presents the results of Fama-Macbeth regressions. Specification 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 Specification 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. Specifications 3, 4, and 5 control for other common factors, while the cyber risk premium remains economically and statistically significant. The economic interpretation of this table is that a one-standard-deviation increase in cyber risk drives returns up by about 18 basis points per month. This increase is statistically significant at the 10 or 5% level, even when controlling for other common factors.

[17] in Appendix, Figure A4 we depict the 2-year rolling window correlation of the cyber risk long-short portfolio and the market. On average, this correlation shows up at only 2%.

[Insert Table 6 here.]

We further test whether an industry effect drives the cyber risk premium. We run Fama-Macbeth regressions, expanding the set of explanatory variables with betas on seven value-weighted Fama-French industry-portfolios[18]. We report these results in Specification 6 of Table 6. The cyber risk premium decreases to 0.150% but remains statistically significant at the 10% level. These 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. We report the results in Specification 7 of Table 6. Once again, the economic significance of the cyber risk premium is virtually unchanged at 0.176% per month, and its statistical significance remains below the 5% level.

4.4.2. GRS test

We implement the GRS test as follows. We build 20 value-weighted portfolios sorted on cyber risk. Next, we compute the GRS test statistic using the five-factor model from Fama and French (2015) and the same model augmented with 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.]

4.4.3. Bayesian factor model selection

Given the results of the GRS tests, a subset of factors could explain the cross-section of returns. The approach of Barillas and Shanken (2018) detailed in Section 3.5 allows us to

[18]We drop four of them due to collinearity and leave the "Other" categories aside because of the lack of interpretability.

determine the optimal combination of factors. Including the market factor in all cases, we compute the 62 permutations between all factors, *i.e.* HML, SMB, CMA, RMW, MOM, and the cyber factor.[19] Figure 6 presents the posterior probabilities of the five most likely models ranked at the end of the sample. All five optimal models contain the cyber risk factor, and the model with the highest probability, of 21.18%, comprises the market, book-to-market, investment, operating profitability, and cyber risk factors.

[Insert Figure 6 here.]

Figure 7 presents the cumulative factor probabilities, *i.e.* 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 high cumulative probabilities.

[Insert Figure 7 here.]

Finally, we study the sensitivity of the model probabilities to the prior multiple of the Sharpe ratio. Following Barillas and Shanken (2018), we repeat the analysis using three other values of prior multiple: 1.25, 2, and 3. Table 8 reports the posterior model probabilities at the end of the sample for the top five models for each prior. The set of top five models is unchanged for each prior multiple. Furthermore, the book-to-market factor is no longer included in the most likely model for higher values of the prior multiple.

[Insert Table 8 here.]

4.5. Comparison with Florackis et al. (2023)

To compare our cyber risk score to the one of Florackis et al. (2023), we start by computing the correlation between the two and obtain 0.34.[20][21] Encouragingly, the correlation is positive but relatively low, which shows that our estimate is novel.

The methodology of this paper improves that of Florackis et al. (2023). Specifically, the Paragraph Vector algorithm is more potent than a dictionary approach in capturing

[19] There are $2^6 - 2$ permutations since we discard the case when all six factors are included in the model.

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

[21] Jamilov et al. (2023) report a correlation coefficient between their measure and that of Florackis et al. (2023) of only 0.39, but with their measure constructed on earning calls.

cyber risk-related content in 10-K statements. Indeed, the dictionary will only contain some words needed to identify all cyber risk-related text. The dictionary used by Florackis et al. (2023) can locate many of these texts unless the authors of the 10-K statements use slightly different wording than the others, which is not uncommon. For example, Appendix Table A3 shows three paragraphs that were missed by their approach, and consequently, the firm-year observations got a cyber risk score of zero. Using the methodology presented in this paper, these 10-K statements were given an above-average cyber risk score.

To investigate this further, we use our cyber risk score to sort firms attributed with a zero cyber risk score by Florackis et al. (2023)) into three portfolios. We report the portfolios' average excess returns and alphas in Table 9. We observe that there is a structure left to exploit in returns sorted from zero-risk firms with dictionary methods. 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 9 here.]

An additional advantage of systematically capturing a score is that it resolves the firm number of the unbalanced portfolios that their method implied, thereby allowing the formation of as many portfolios as needed. Finally, we test how much cyber risk information is missed in 10-K statement parts other than Items 1. A. We sort portfolios based on the cyber score that excludes this section, and we report the results in Table 10. We observe that alphas and statistical significance levels are slightly lower, but our interpretation is unchanged. Thus, we conclude that the entirety of the filings needs to be considered since a significant amount of cyber risk-related information remains outside of Item 1. A. Specifically, computing other statistics, we find that, on average, only 34% of paragraphs used for our cyber risk score come from Items 1. A. This proportion has increased regularly, from about 15% in 2008 to about 50% in 2022. We also report an excerpt of these Paragraph Vector-identified top-scoring paragraphs in Appendix Table A3.

[Insert Table 10 here.]

In unreported tests, we also circumvent the cyber risk score to be only on Item 1. A: We obtain slightly less significant results, further supporting the idea that the information should be extracted from the full statement.

4.6. Robustness tests

4.6.1. Long-run cyber risk

By design, our cyber risk score depends only on each company's most recent 10-K statement. A firm may discuss 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-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 score would miss these cases as it has no memory.

We compute the expanding average cyber risk score to study firms' long-run cyber risk. 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 estimate could account for the case 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. Thus, a low cyber risk score in the subsequent years would not represent the firm's reality.

We repeat the portfolio sorts from Section 4.2 using the long-run cyber risk to sort firms. Table 11 presents the results. We observe no significant change from Table 4, and the long-short portfolio remains significant at the 5% or 10% level. These results indicate that firms incorporate all available information regarding their cyber risk in their newest 10-K statements; hence, incorporating information from past statements does not improve the estimation of the cyber risk.

[Insert Table 11 here.]

4.6.2. Controlling for cybersecurity firms

Our cyber risk score 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[22]. As there is no dedicated cybersecurity industry classification, we identify cybersecurity firms using the HACK ETF[23]. The fund's description explains that this ETF invests in companies providing cybersecurity solutions, including hardware, software, and services. As cybersecurity providers, these firms are expected to discuss cybersecurity extensively in their 10-K statements, 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 12 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.

[Insert Table 12 here.]

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

We repeat our analysis by splitting the study period before and after 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 13 presents the results using the period before the publication. 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.

[Insert Table 13 here.]

Table 14 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

[22]Available at: https://www.fortinet.com/
[23]Available at: https://etfmg.com/funds/hack/

are not statistically significant when controlling for the factors from Carhart (1997) or Fama and French (2015). Last, the correlation between the long-short cyber risk and the market portfolios spikes in this period (see Appendix, Figure A4).

[Insert Table 14 here.]

There could be several explanations for these results. It could be that the publication made some arbitrageurs trade stocks based on their cyber risk evaluation, which results in lower returns post-publication, as explained in McLean and Pontiff (2016). Another interpretation provided by Florackis et al. (2023) is that high-risk firms lost value in a higher concentration of cyber incidents while low-risk firms, being less affected, 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[24]. 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[25]. The report also found that the cost of ransomware payments increased. These events could explain why Portfolio 5 has low returns, and the long-short portfolio has negative returns. However, the post-publication period is much smaller; thus, the estimates could be spurious.

5. Conclusion

This paper implements the "Paragraph Vector" algorithm to estimate firms' cyber risk based on their 10-K statements. Documenting that this score is unrelated to other economic variables, we sort portfolios across this cyber risk score that we submit to 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 over traditional factor models and an average monthly excess return of 0.56%. We investigate the interactions of cyber risk with other sources of risk 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

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

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 compare our score to the one of Florackis et al. (2023) and conclude that while being mildly correlated with theirs (34%), our score performs better at capturing cyber risk. By design, our method allows us to capture cyber risk contexts and not only specific keywords. Consequently, this method can be unrestricted to specific sections of firms' disclosures, thereby integrating all available information. We confirm this with a specific test on firms sorted on 10-K statements, excluding Item 1. A and document excess returns in line with those obtained using the full corpus. Second, our approach always enables us to obtain a firm cyber risk score, avoiding the issue of classifying firms as zero-risk. Specifically, while sorting firms attributed zero-cyber risk with the dictionary approach, we uncover a remaining structure in the portfolio's returns with long-short portfolio returns statistically significant at the 5% level and of 0.53% per month on average. A third benefit of this approach is that it does not require the use of past cyberattacks to be calibrated, which solves the problem of whether what is captured are "post-cyberattacks recovery", or actual latent cyber risk.

Finally, we run two robustness tests. First, we 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. Second, 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 arises from cyber-risk exposure. We are confident that this new semantic approach incorporating information in textual documents would successfully identify other risk factors or characteristics, such as legislative or geopolitical risks. 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.

Amir, E., Levi, S., Livne, T., 2018. Do firms underreport information on cyber-attacks? Evidence from capital markets. Review of Accounting Studies 23, 1177–1206.

Anderson, R., Barton, C., Böhme, R., Clayton, R., Eeten *van*, 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.

Anderson, R., Barton, C., Böhme, 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.

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`

Antweiler, W., Frank, M. Z., 2004. Is all that talk just noise? The information content of internet stock message boards. Journal of Finance 59, 1259–1294.

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.

Bodnaruk, A., Loughran, T., McDonald, B., 2015. Using 10-K text to gauge financial constraints. Journal of Financial and Quantitative Analysis 50, 623–646.

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

Calomiris, C. W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. Journal of Financial Economics 133, 299–336.

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. Journal of Finance 52, 57–82.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. Journal of Finance 75, 1371–1415.

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. Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1997. Industry costs of equity. Journal of Financial Economics 43, 153–193.

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

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

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

Garcia, D., 2013. Sentiment during recessions. Journal of Finance 68, 1267–1300.

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.

Hassan, T. A., Hollander, S., *van* Lent, L., Tahoun, A., 2019. Firm-level political risk: Measurement and effects. Quarterly Journal of Economics 134, 2135–2202.

Hoberg, G., Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. Review of Financial Studies 23, 3773–3811.

Jamilov, R., Rey, H., Tahoun, A., 2023. The anatomy of cyber risk. Available at: https://ssrn.com/abstract=3866338

Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729.

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 16, 1–13.

Kamiya, S., Kang, J.-K., Kim, J., 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

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

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

Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., Shimada, H., 2019. Identification of cybersecurity specific content using the doc2vec language model. In: *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, pp. 396–401.

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.

Ross, S. A., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360.

Sautner, Z., *van* Lent, L., Vilkov, G., Zhang, R., 2023. Firm-level climate change exposure. Journal of Finance 78, 1449–1498.

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

Figures and Tables

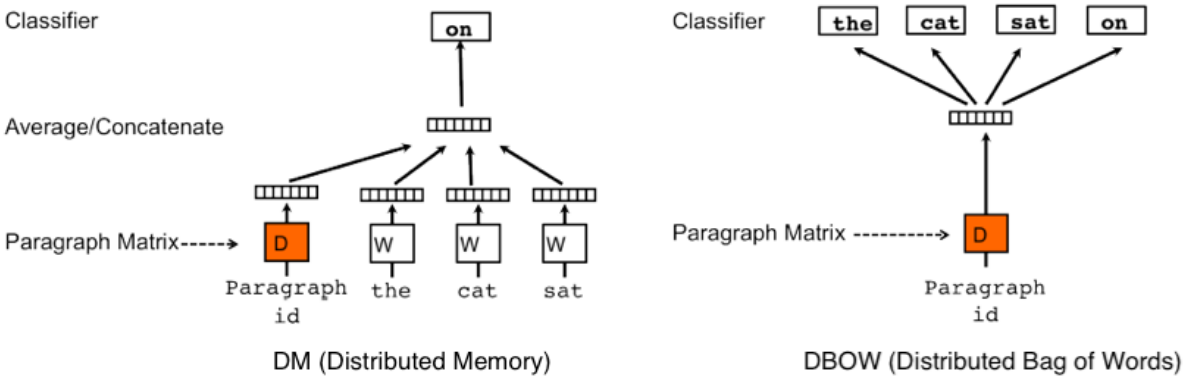


Fig. 1: **"Paragraph Vector" model versions**

The left panel of the figure depicts the framework for learning paragraph vectors. A paragraph token is mapped to a vector via matrix D (in orange). In this case, the concatenation of this vector with a context of three words is used to predict the fourth word ("on"). The paragraph vector represents the missing information from the context and is a memory of the paragraph's topic. The right panel depicts the "Distributed Bag of Words" version of paragraph vectors. This version trains the paragraph vector to predict the words in a small window. The images (captions) are taken (rewritten) from Le and Mikolov (2014).

	Method	Training Size	Vector Size	Window Size	Min Count	Sub-Sampling	Negative Sampling	Epoch
Best specification	DBOW	1.7M	200	15	5	10^{-5}	5	50
Baseline specification	DBOW	-	300	15	5	10^{-5}	5	20

Table 1: **doc2vec parameters**

This table reports the chosen parameters used to train the doc2vec specification and the baseline specification parameters of Lau and Baldwin (2016). The chosen parameters correspond to the best specification in computing the vector representations of the paragraphs from 10 randomly selected 10-K statements from 2008 (validation sample). We try each specification and compare the highest-scoring paragraphs between the specifications. We assume the best specification to be the one with the highest proportion of highest-scoring cyber risk-related paragraphs. "DBOW" stands for distributed bag-of-words.

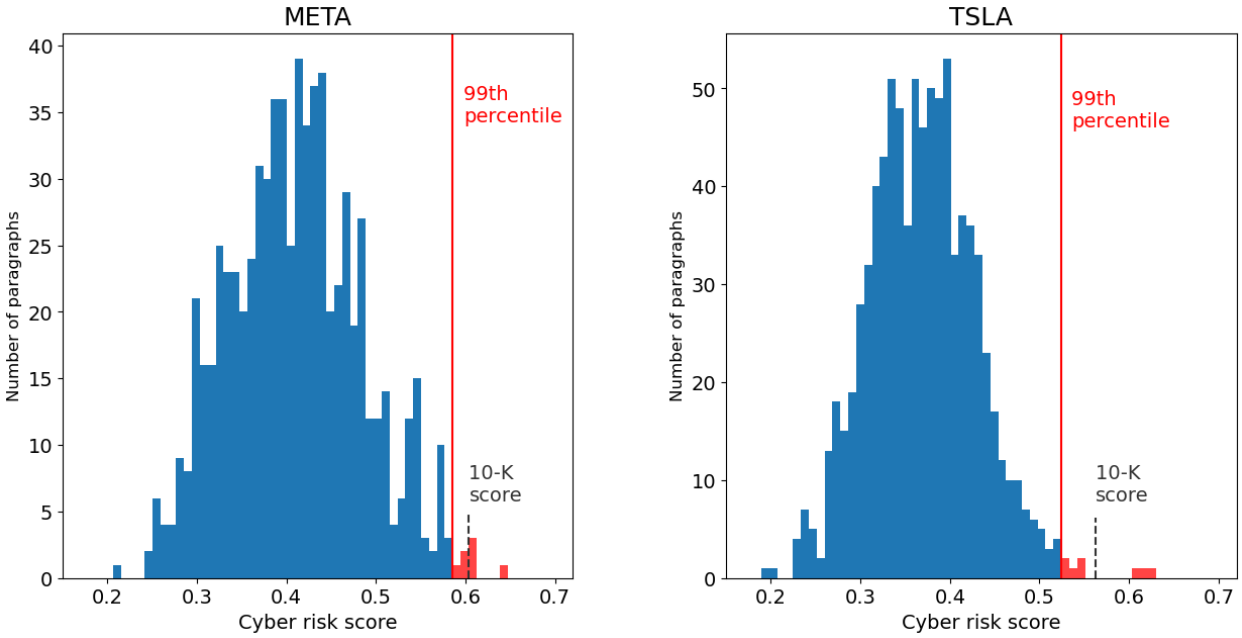


Fig. 2: **Paragraph level score distributions for Meta Platforms and Tesla**

This figure depicts the distribution of paragraphs from 10-K statements filed in 2022 for Meta (left panel) and Tesla (right panel) based on their cyber risk score. We choose 50 bins and color in red the paragraphs within the top 1% of cyber risk scores.

	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
Readability	16.02	1.10	13.12	15.40	16.25	16.78	18.03	0.11
Risk section length	4.98	1.46	0.00	4.80	5.30	5.76	6.79	0.21
Turnover	0.28	6.52	0.00	0.06	0.13	0.23	1.97	0.02

Table 2: **Descriptive statistics of the cyber risk measure and firm characteristics**

This table reports firm-level characteristics, Mean, standard deviation, percentiles, and correlation with cyber-risk. The measure is winsorized at the 1^{st} and 99^{th} percentile (by year). We define the characteristics in Table A1.

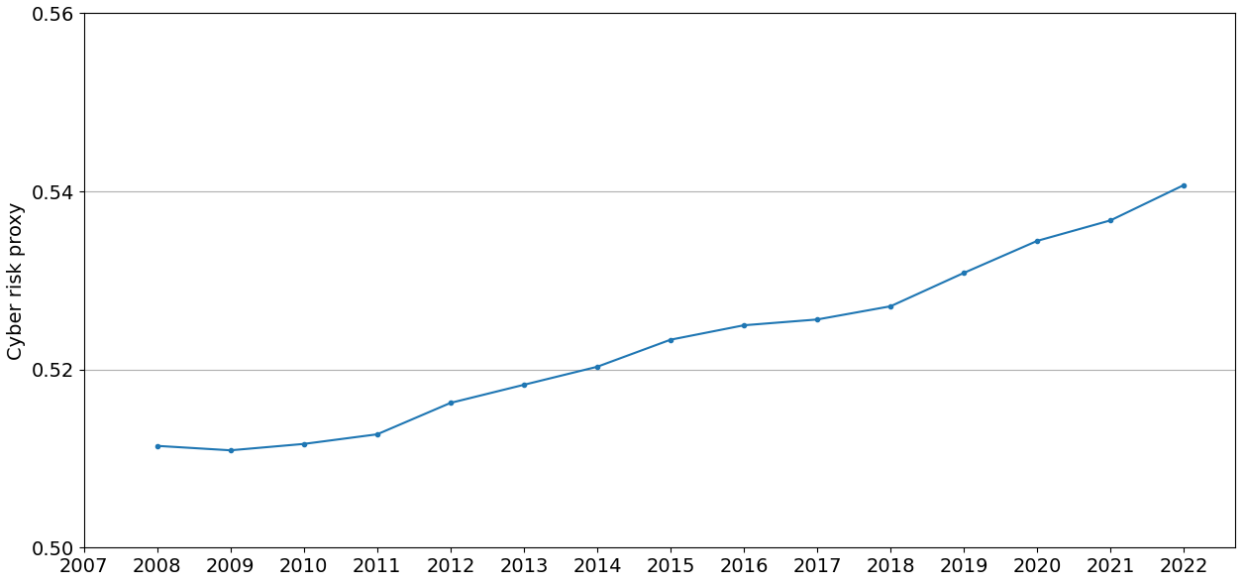


Fig. 3: **Evolution of the average cyber risk across all firms**

This figure depicts the yearly evolution of our firm-level cyber risk measure averaged across all firms over our sample period: 2007–2022.

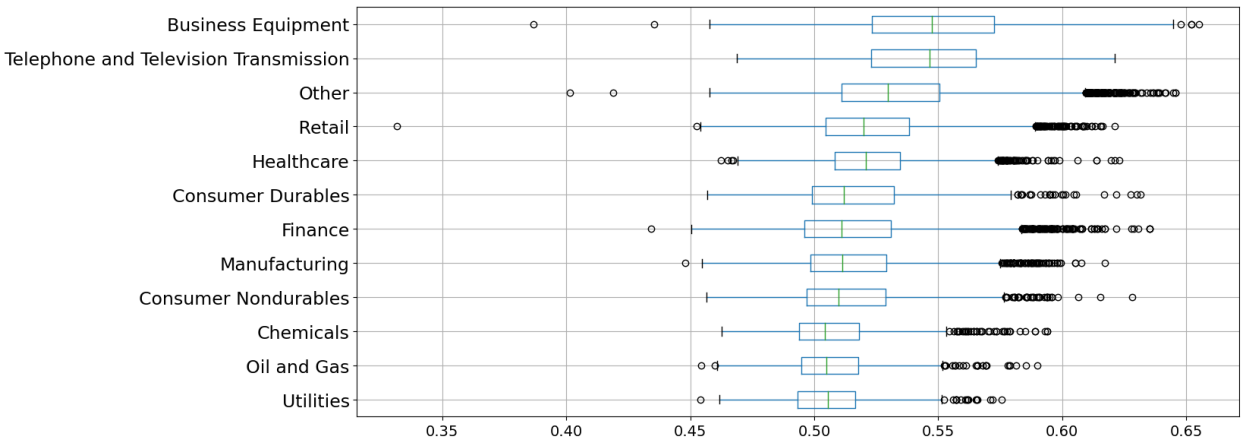


Fig. 4: **Cyber risk across industries**

This box plot depicts the minimum, first quartile, median, third quartile, and maximum values of our cyber risk measure for each Fama-French 12 industry classification. The whiskers extend no more than 1.5 times the interquartile range (third quantile - first quantile) from the edges of the box, ending at the farthest data point within that interval. Outliers are plotted as separate dots. 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.

Firm-level cyber risk score		
	Model 1	Model 2
Constant	**0.50*****	**0.53*****
	[45.22]	[64.17]
Firm Size (ln)	**0.02**	**0.03**
	[0.33]	[1.05]
Firm Age (ln)	**-0.33*****	**-0.56*****
	[-2.70]	[-9.68]
ROA	**-0.00**	**0.00**
	[-0.03]	[0.01]
Book to Market	**-0.04*****	**-0.18*****
	[-3.44]	[-5.39]
Tobin's Q	**0.03*****	**0.17*****
	[2.62]	[8.10]
Market Beta	**-0.06***	**-0.11*****
	[-1.88]	[-2.58]
Intangibles/Assets	**-0.29**	**1.21*****
	[-1.55]	[5.63]
Debt/Assets	**-0.42***	**1.11****
	[-1.81]	[2.54]
ROE	**0.01**	**-0.03**
	[0.21]	[-0.28]
Price/Earnings	**0.00****	**-0.00**
	[2.13]	[-0.26]
Profit Margin	**0.02**	**0.03*****
	[0.37]	[2.94]
Asset Turnover	**-0.05**	**-0.49*****
	[-0.57]	[-3.80]
Cash Ratio	**0.00**	**0.02**
	[0.24]	[0.10]
Sales/Invested Capital	**0.08**	**0.14****
	[0.23]	[2.21]
Capital Ratio	**0.00**	**-2.40*****
	[0.00]	[-7.15]
R&D/Sales	**-0.00**	**0.00**
	[-0.62]	[0.33]
ROCE	**0.04**	**0.31***
	[0.41]	[1.80]
Readability (ln)	**0.09*****	**-0.14****
	[2.84]	[-2.51]
Risk section length (ln)	**0.13*****	**0.28*****
	[4.38]	[7.94]
Turnover	**-0.01**	**0.00*****
	[-1.44]	[3.74]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	23617	23617
R^2_{within}	0.2958	0.2385

Table 3: **Determinants of firm-level cyber risk**

This table reports the results of cyber risk regressions on firm characteristics. We include in turn year-, industry-, and firm-fixed effects. We report t-statistics in brackets. The variables are standardized, and the standard errors are clustered at the firm level. The reported R^2_{within} is the within estimator. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. We define the characteristics in Table A1.

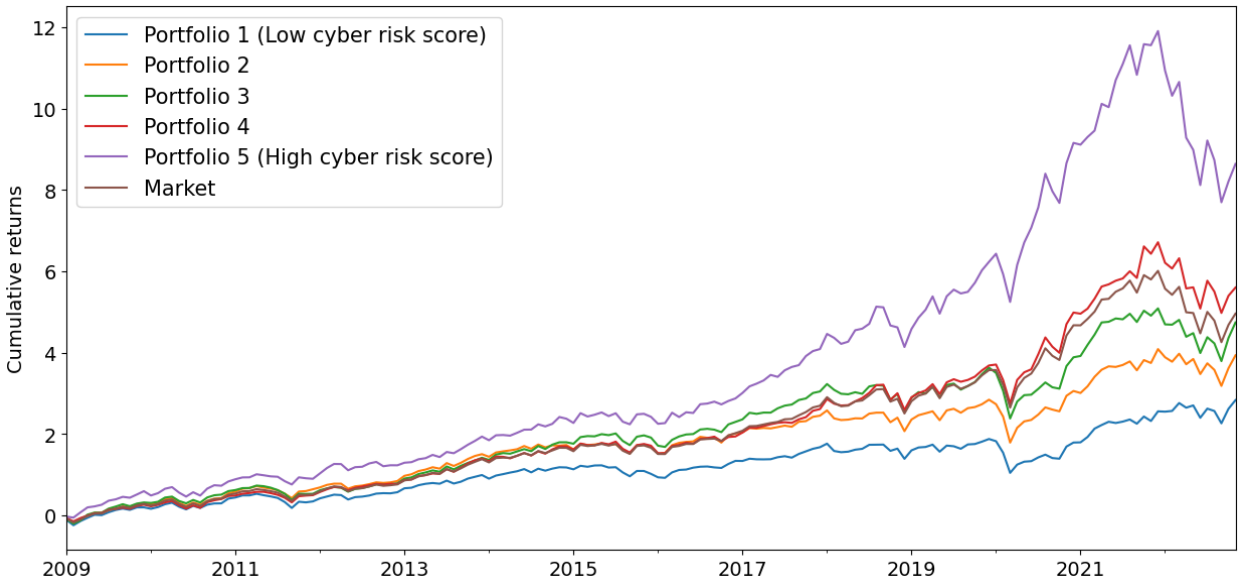


Fig. 5: **Cyber risk-sorted portfolios cumulative returns**

This figure depicts the performance of value-weighted portfolios of firms sorted on our cyber risk measure. We rebalance the portfolios quarterly. "Market" refers to the market portfolio obtained from the Kenneth French data repository. One unit of the ordinate axis unit is 100% return. The frequency is monthly, and the period is from 2009 to 2022.

	Value-weighted portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber risk						
Average excess returns	**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 4: **Average monthly excess returns and alphas**

This table reports the excess returns and alphas (in percent) of quintile portfolios of firms sorted on our cyber risk measure. "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 reports the average number of firms in each portfolio, the average cyber risk, and the annualized Sharpe, Treynor, and Sortino ratios. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to December 2022.

	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 5: **Double sorted portfolios**

This table reports average monthly excess returns in percent of double-sorted portfolios, first on market beta, book-to-market, or size and then on cyber risk. The period is from January 2009 to December 2022.

Portfolio returns $R_{P,t+1}$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\beta_{Market,t}$	**-0.005**		**-0.025**	**0.065**	**0.024**		**-0.033**
	[-0.064]		[-0.429]	[0.997]	[0.275]		[-0.578]
Cyber risk$_t$		**0.183***	**0.182****	**0.183***	**0.172****	**0.150***	**0.176****
		[1.794]	[1.994]	[1.913]	[2.037]	[1.736]	[2.103]
$\beta_{HML,t}$				**0.027**	**-0.012**		**-0.033**
				[0.439]	[-0.126]		[-0.293]
$\beta_{SMB,t}$				**0.069**	**0.049**		**-0.042**
				[0.835]	[0.536]		[-0.385]
$\beta_{MOM,t}$				**0.011**			
				[0.153]			
$\beta_{RMW,t}$					**-0.085**		**-0.086**
					[-1.436]		[-1.150]
$\beta_{CMA,t}$					**-0.088**		**0.011**
					[-0.816]		[0.076]
$\beta_{ConsumerDurables,t}$						**-0.005**	
						[-0.053]	
$\beta_{Energy,t}$						**-0.005**	
						[-0.064]	
$\beta_{Chemicals,t}$						**0.025**	
						[0.351]	
$\beta_{Telecommunications,t}$						**0.045**	
						[0.619]	
$\beta_{Retail,t}$						**0.116**	
						[1.015]	
$\beta_{Healthcare,t}$						**0.048**	
						[0.623]	
$\beta_{Finance,t}$						**0.101**	
						[1.104]	
Constant	**1.445*****	**1.465*****	**1.457*****	**1.455*****	**1.476*****	**0.945****	**1.362*****
	[5.311]	[5.540]	[5.493]	[5.413]	[5.450]	[2.525]	[5.015]
$\overline{R^2_{adj}}$	0.007	0.134	0.186	0.258	0.284	0.302	0.295
MAPE	1.360	1.312	1.233	1.064	0.987	0.914	1.243
Nb portfolios	20	20	20	20	20	20	31

Table 6: **Fama-MacBeth regressions**

This table reports the results of Fama-MacBeth regressions of 20 value-weighted portfolios arithmetic returns sorted on their cyber risk. In Specification (7), we extend the set of test assets with an additional 11 industry portfolios. These portfolios are regressed each month on portfolio value-weighted betas with the market, HML, SMB, MOM, RMW, and CMA and on portfolio value-weighted betas with industry portfolios in Specification (6). "Cyber risk" is the value-weighted cyber-risk of each portfolio. The betas are standardized before the second step-regressions and estimated with a two-year rolling window. 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). $\overline{R^2_{adj}}$ 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. The period is from January 2009 to December 2022.

	GRS	p-value	$\overline{R^2}$	GRS	p-value	$\overline{R^2}$
	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**

This table reports the results of time series regressions of 20 value-weighted portfolios sorted on cyber-risk on the five-factor model of Fama and French (2015) (FF5) and on the "CyberFactor", *i.e.* the factor built as the long-short of extreme quintile portfolios sorted on our cyber risk score. The study period is from January 2009 to December 2022.

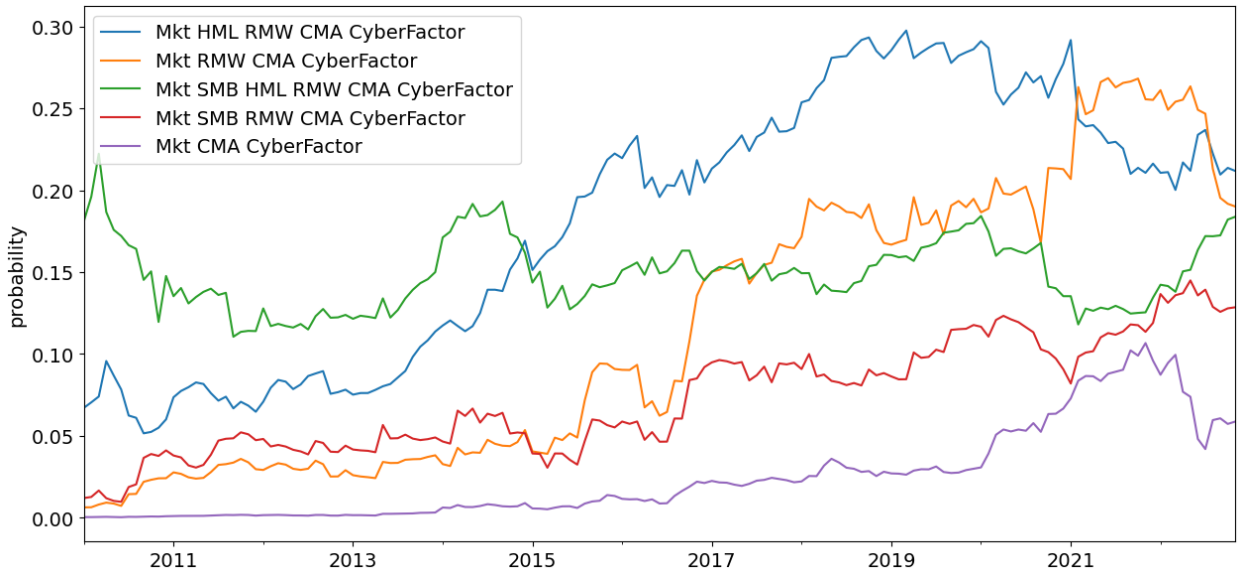


Fig. 6: **Factor model posterior probabilities**

The figure depicts the posterior probabilities for the top five models, ranked at the end of the sample. "Mkt" refers to the excess return on the market from the Kenneth French data repository. HML and SMB refer to the book-to-market and size factors of Fama and French (1992). CMA and RMW refer to the investment and operating profitability factors of Fama and French (2015). "CyberFactor" refers to the long-short portfolio built on our cyber risk score. The prior multiple is 1.5, and the study period is from January 2010 to December 2022.

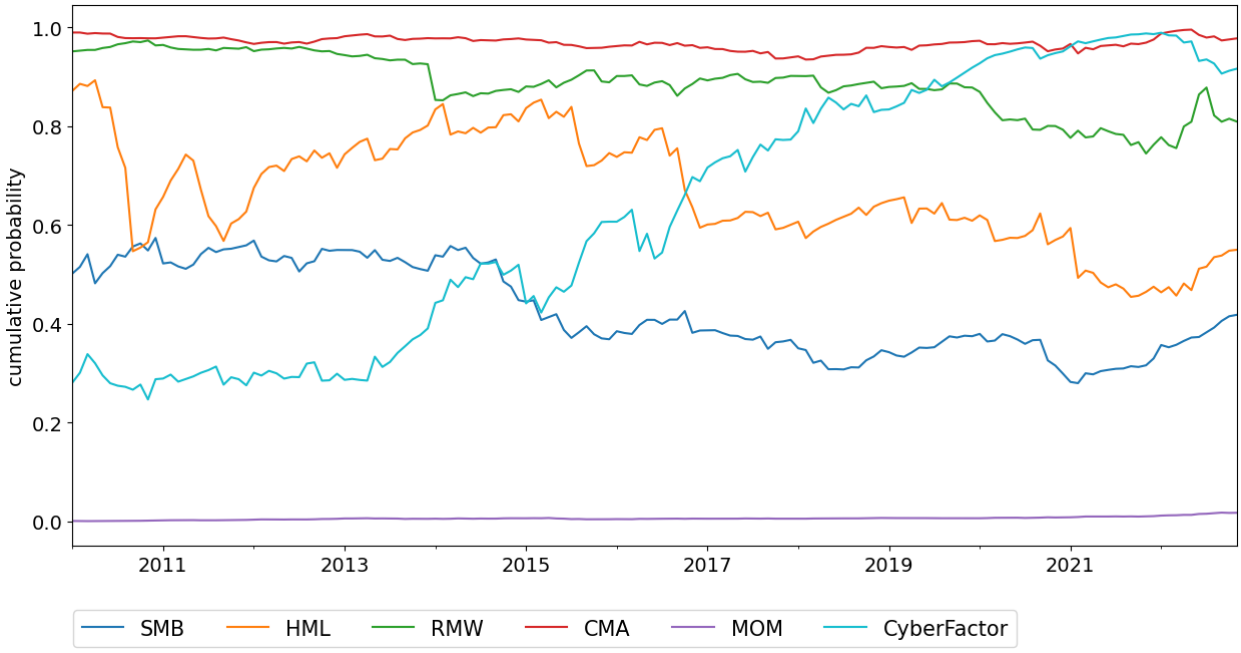


Fig. 7: **Cumulative posterior factor probabilities**

This figure depicts the cumulative posterior probabilities, *i.e.* 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 built on our cyber risk score. The prior multiple is 1.5, and the study period is from January 2010 to December 2022.

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**

This table reports the posterior model probabilities (in percent) for the top five models (ranked at the end of the sample) for different values of the prior multiple. The top five models remain the same for every prior multiple. "Mkt" refers to the excess return on the market from the Kenneth French data repository. HML and SMB refer to the book-to-market and size factors of Fama and French (1992). CMA and RMW refer to the investment and operating profitability factors of Fama and French (2015). "CyberFactor" refers to the long-short portfolio built on our cyber risk score. The study period is from January 2010 to December 2022.

	Value-weighted portfolios			
	L		H	H-L
	P1	P2	P3	P3-P1
A. Portfolios sorted by cyber risk				
Average excess returns	**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	247.9	248.3	-
Cyber risk	0.486	0.504	0.534	-

Table 9: **Average monthly excess returns and alphas of firms classified as zero-risk by Florackis et al. (2023)**

This table reports the average monthly excess returns and alphas (in percent) of firms classified as zero-risk when using the procedure 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 reports the average number of firms in each portfolio and the average cyber risk. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to December 2018.

	Value-weighted portfolios					
	L P1	P2	P3	P4	H P5	H-L P5-P1
A. Portfolios sorted by cyber risk						
Average excess returns	**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	-

Table 10: **Average monthly excess returns and alphas omitting Items 1A**

This table reports the average monthly excess returns and alphas (in percent) when items 1A from 10-K statements are omitted. 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 reports the average number of firms in each portfolio and the average cyber risk. Newey-West (Newey and West, 1994) t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to December 2022.

	Value-weighted portfolios					
	L				H	H-L
	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by the long-run cyber risk						
Average excess returns	**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	-

Table 11: **Average monthly excess returns and alphas using the long-run cyber risk**

This table reports the average monthly excess returns and alphas (in percent) when the portfolios are sorted on our long-run cyber risk measure, *i.e.* the cumulative cyber risk for each firm over the period. FFC refers to the four-factor model of Carhart (1997), and FF5 refers to the five-factor model of Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average long-run cyber risk. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to December 2022.

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

Table 12: **Average monthly excess returns and alphas excluding cybersecurity firms**

This table reports the average monthly excess returns and alphas (in percent) when we exclude cybersecurity firms from the sample. We identify cybersecurity firms from the "HACK" cybersecurity ETF. FFC refers to the four-factor model of Carhart (1997), and FF5 refers to the five-factor model of Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to December 2022.

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

Table 13: **Average monthly excess returns and alphas before the first release of Florackis et al. (2023).**

This table reports the average monthly excess returns and alphas (in percent) before the first release of Florackis et al. (2023) on SSRN in November 2020. 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 reports the average number of firms in each portfolio and the average cyber risk. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from January 2009 to October 2020.

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

Table 14: **Average monthly excess returns and alphas after the first release of Florackis et al. (2023).**

This table reports the average monthly excess returns and alphas (in percent) after the first release of Florackis et al. (2023) on SSRN in November 2020. 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 reports the average number of firms in each portfolio and the average cyber risk. Newey-West t-statistics are reported in brackets. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. The period is from November 2020 to December 2022.

Appendix

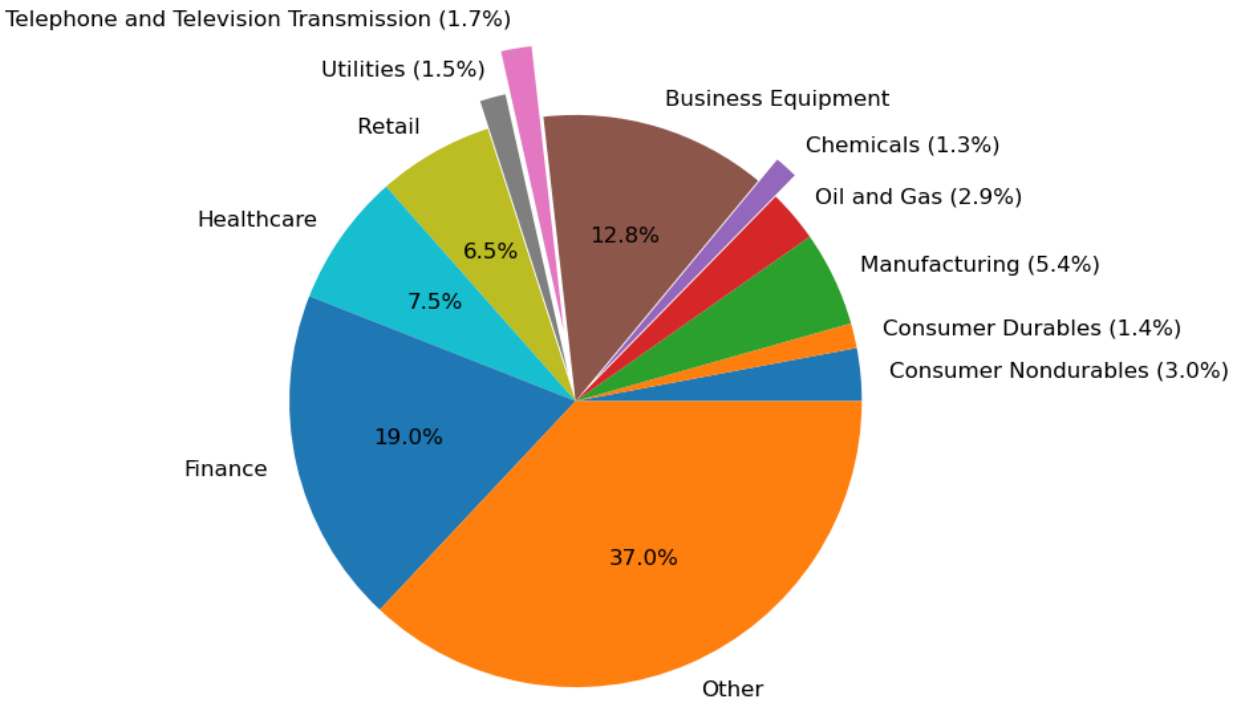


Fig. A1: **Industry distribution**

This figure depicts the 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.

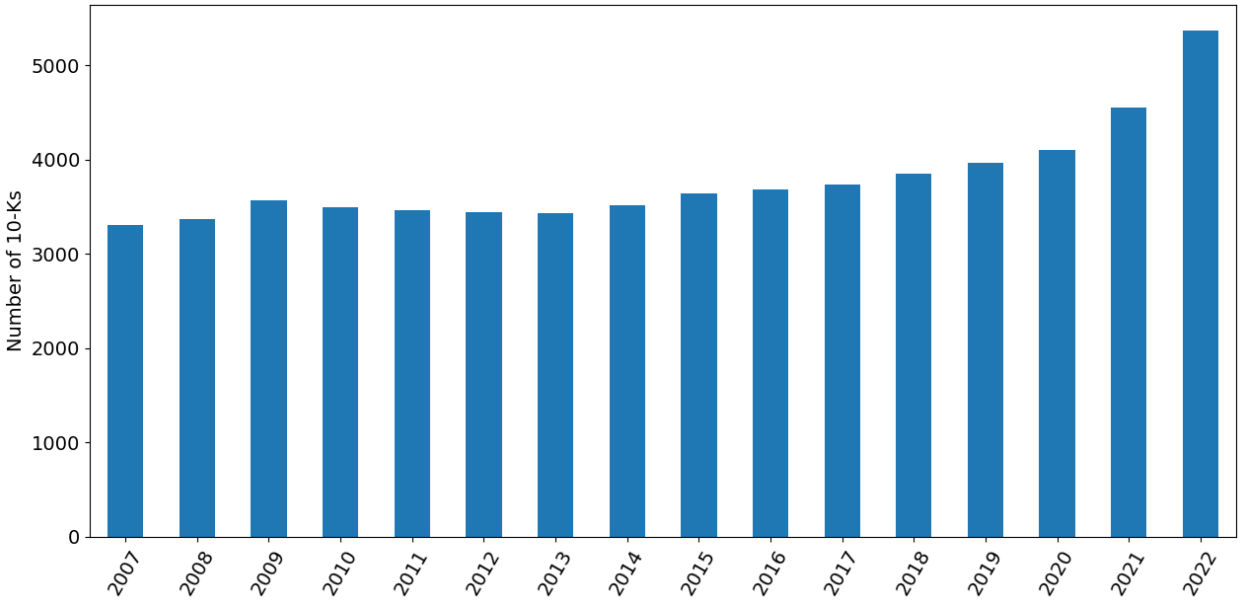


Fig. A2: **Number of 10-Ks per year**

This figure depicts the number of companies in the study sample that have filed a 10-K statement in given calendar year.

Panel A.

Tactic

Tactics represent the "why" of an ATT&CK technique or sub-technique. It is the adversary's tactical goal: the reason for acting. 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 acting. 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 techniques. 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.

Panel B.

		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	

Fig. A3: **Structure and example of the tactic descriptions of MITRE ATT&CK**

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
Readability (ln)	Number of characters in the 10-K	EDGAR - SEC
Risk section length (ln)	Number of sentences in Item 1A of the 10-K	EDGAR - SEC
Turnover	Monthly trading volume / Market capitalization	CRSP

Table A1: **Variable definitions**

This table reports the variable names used throughout the paper, their description, and their source. Square brackets indicate variable name definitions in CRSP and Compustat.

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
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

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

This exhibit reports the paragraphs selected 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 ten companies in the validation sample are CSCD, GTS, LXRX, MDAS, PBY, PZZA, UMH, VALU, VSTY and VXRT.

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 A3: **Examples of paragraphs missed by the dictionary approach of Florackis et al. (2023)**

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 A4: **Examples of negative scoring paragraphs**

This exhibit reports 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. Given their low cyber risk score, we select two firms, AGCO Corporation (AGCO) and O-I Glass Inc (OI). The 10-K statements are filed in 2022.

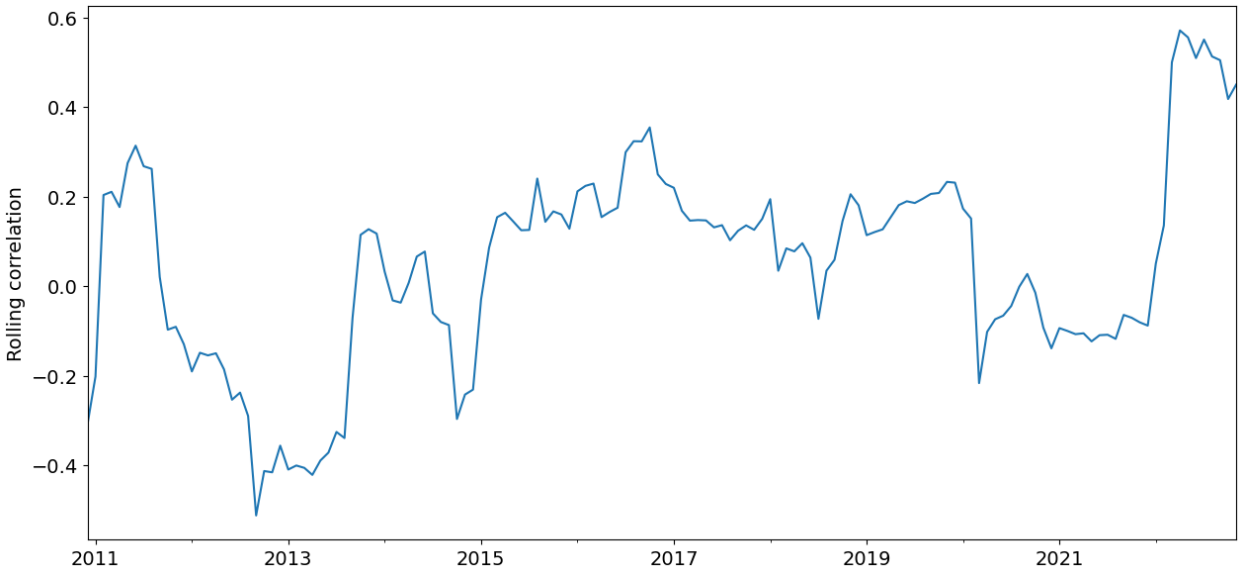


Fig. A4: **Rolling correlation between the returns on the long-short cyber risk portfolio and the market**

This figure depicts the rolling correlation between the monthly returns on the long-short portfolio built on the cyber risk score and the market portfolio. The rolling window is two years, and the period is from January 2011 to December 2022.