Disentangling the sources of cyber risk premia

Loïc Maréchal* and Nathan Monnet†

Abstract

We use a methodology based on a machine learning algorithm to quantify firms' cyber risks based on their disclosures and a dedicated cyber corpus. The model can identify paragraphs related to determined cyber-threat types and accordingly attribute several related cyber scores to the firm. The cyber scores are unrelated to other firms' characteristics. Stocks with high cyber scores significantly outperform other stocks. The long-short cyber risk factors exhibit positive risk premia, are robust to all factors' benchmarks, and help price returns. Furthermore, we suggest the market does not distinguish between different types of cyber risks but instead views them as a single, aggregate cyber risk. Importantly, our cyber score captures the extent to which firms' disclosures are semantically related to known cyber attack descriptions and therefore reflects exposure, awareness, or discussion of cyber risk, rather than realized cyber incidents. We provide evidence that the score is predictive of future cyber-related disclosures, supporting its interpretation as a proxy for latent cyber risk.

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

*HES-SO Valais-Wallis, Institute of Entrepreneurship & Management. loic.marechal@hevs.ch

†Swiss Finance Institute, École Polytechnique Fédérale de Lausanne - Cyber-Defence Campus, armasuisse S+T. nathan.monnet@armasuisse.ch

This document results from a research project funded by the Cyber-Defence Campus, armasuisse Science and Technology, and was initially written as Nathan Monnet's Master thesis. We appreciate helpful comments from seminar participants at the Cyber Alp Retreat 2024, MRS Conference 2025, and Modern Finance Conference 2025. We also thank Julien Hugonnier and Michel Dubois for their invaluable comments. Corresponding author: Loïc Maréchal e-mail: loic.marechal@hevs.ch

1. Introduction

Cyber-insurance contracts and cybersecurity solutions have become crucial for private and public organizations in the widespread and costly context of cyber incidents. These countermeasures, however, have costs that are challenging to estimate. This paper aims to overcome this challenge by utilizing natural language processing, clustering methods, and state-of-the-art asset pricing techniques to disentangle and quantify the risk premia associated with various cyber threats.

A key challenge in measuring cyber risk using textual disclosures is interpreting what these measures capture. Mentions of cyber-related terminology in 10-K filings may reflect realized incidents, anticipated risks, regulatory disclosure practices, or general awareness of cybersecurity issues. In this paper, we interpret our cyber score as a measure of textual exposure to cyber risk rather than a direct measure of realized cyber incidents. This distinction is important for interpreting the empirical results and aligns with prior literature using disclosure-based risk measures.

This paper contributes to the literature in three main ways. First, we extend prior work by decomposing cyber risk into distinct categories derived from the MITRE ATT&CK framework, allowing for a more granular analysis of cyber threats. Second, we test whether financial markets differentiate between these sources of cyber risk and find that, despite their heterogeneity, markets price them as a single aggregate risk factor. Third, we provide additional validation of our cyber score by showing that it predicts future cyber-related incidents disclosed in 8-K filings.

To do this, we collect financial filings, monthly returns, and other firm characteristics for over 7000 firms listed on US stock markets between January 2007 and December 2023. We use a neural network called "Paragraph Vector" in combination with the MITRE ATT&CK cybersecurity knowledgebase and clustering techniques to score each firm's filing based on its various types of cyber risk.

We identify four types of cyberattacks that emerge from the textual cluster structures of MITRE ATT&CK. We establish scores to quantify the similarity between the annual statements of firms, the 10-Ks, and the identified types of cyber attacks from the MITRE ATT&CK knowledge base. We find that the four cyber scores present no correlation with standard firms' characteristics known to influence stock returns and weak correlations with

textual non-semantic variables of the annual statement (the highest correlation, 0.36, is with the length of section 1.A. in the 10-Ks). As previously observed in Celeny and Maréchal (2023), who use the same neural network, the resulting aggregation of the various cyber scores shows an increasing trend, with scores increasing by 0.04 from 2007 to 2023. Additionally, specific industries from the Fama-French 12-industry classification exhibit higher cyber scores, with Business Equipment and Telephone and Television Transmission scoring the highest.

Sorting firms into cyber scores-based portfolios, we observe monotonically increasing average excess returns along the sorts. All average excess returns of the portfolios are statistically significant at the 1% level, and investing in a portfolio that enters a long (short) position in the top (bottom) cyber scoring firm is statistically significant at the 5% level. After controlling for common risk factors, the results above remain valid at the 5% and 10% levels in the top portfolios.

The risk premia associated with the different types of cyber risk are also significant at the 5% level in the cross-section with Fama and MacBeth (1973) regressions. Using additional pricing factors related to cyber-based portfolios improves pricing ability. We demonstrate that joint alphas of various assets tend to decrease in Gibbons, Ross, and Shanken (1989) tests. Using the Bayesian approach of Barillas and Shanken (2018), we also show that the optimal subset of factors pricing stock returns invariably includes the cyber-based factors.

Additional tests reveal that, although various types of cyber risk exist, the market does not differentiate between them and perceives them as a single aggregate cyber risk. Finally, we conduct an event study to evaluate the performance of a cyber-based portfolio during the massive SolarWinds cyber attack in December 2020. Contrary to previous studies, particularly Florackis, Louca, Michaely, and Weber (2023), no significant conclusions can be drawn from this event regarding the performance of our cyber-based portfolios in response to cyberattacks.

The remainder of this work proceeds as follows. Section 2 introduces the related literature and develops hypotheses. Section 3 presents the data and methodology, Section 4 outlines the results, and Section 5 concludes.

2. Literature review

2.1. Sentiment analysis and text classification

This study connects to several strands of the literature on sentiment analysis and its application to financial markets. First, it relates to the growing body of research that uses textual data to extract economic insights from corporate filings and other financial disclosures (*e.g.*, Antweiler and Frank, 2004; Garcia, 2013; Arslan-Ayaydin, Boudt, and Thewissen, 2016). Early studies, such as Feldman, Govindaraj, Livnat, and Segal (2010), examine the Management Discussion and Analysis (MD&A) sections of 10-Q and 10-K filings and demonstrated that changes in the tone of non-financial information, measured by the frequency of positive and negative words, correlate with both short-term market returns and long-term excess returns. This study highlighted the importance of textual sentiment in SEC filings, laying the groundwork for more nuanced measures of market sentiment. Similarly, Jegadeesh and Wu (2013) introduce a return-based term weighting scheme to extract document tone in 10-K filings. They show that the tone of these documents, especially when weighted using positive and negative dictionaries, effectively predicts market reactions around filing dates. Their method also generalizes well to other financial documents, such as IPO prospectuses, further expanding the applicability of sentiment analysis in financial contexts.

Building on these foundational studies, Antweiler and Frank (2004) apply sentiment analysis to online forums, showing how sentiment derived from financial discussions can predict market returns and trading volumes. Garcia (2013) extends the scope of sentiment analysis to financial news, revealing that the predictive power of sentiment on stock returns is particularly pronounced during recessions. These studies collectively demonstrate that textual sentiment, whether derived from financial disclosures or public discourse, is crucial in predicting market behavior. Furthermore, Bodnaruk, Loughran, and McDonald (2015) advance textual analysis by developing a linguistic measure of firm-level financial constraints based on 10-K disclosures. Their approach, which identifies constraining words such as "required" or "obligations", proved to be more effective than traditional financial constraint indexes in predicting liquidity events, offering an innovative way to assess financial health through text.

This study contributes to the literature on textual analysis to assess specific risks. Hassan, Hollander, *van* Lent, and Tahoun (2019) and Sautner, *van* Lent, Vilkov, and Zhang (2023)

apply sentiment analysis to earnings conference calls to measure firm-level political risk and climate risk, respectively. Hassan et al. (2019) use political language to measure political risk, showing that political discussions during earnings calls significantly influence firm behavior and stock volatility. In contrast, Sautner et al. (2023) focus on climate-related language, using a keyword discovery algorithm to assess the extent of firms' climate risk exposure. These studies demonstrate how sentiment analysis can be tailored to identify specific types of risk beyond traditional financial metrics, thus expanding the scope of risk analysis to include socio-political and environmental factors.

In addition to extending the application of sentiment analysis, Calomiris and Mamaysky (2019) incorporate advanced text-processing techniques, such as word flow measures, to predict market returns and risks across both developed and emerging markets. Analyzing sentiment, frequency, and entropy, they demonstrate that textual data could capture latent market risks more effectively than traditional methods. This approach complements the earlier work of Jegadeesh and Wu (2013) and Bodnaruk et al. (2015) by showing how textual measures can be fine-tuned to assess not just sentiment but also financial constraints and specific risk factors, such as market volatility or liquidity challenges.

This study ties into the broader literature on how textual sentiment influences firm and market behavior. Antweiler and Frank (2004) show that sentiment derived from public discussions could affect stock prices and trading volumes. Arslan-Ayaydin et al. (2016) demonstrate how managerial incentives shape the tone of earnings press releases, influencing market reactions.

2.1.1. Vector representation of paragraphs and topics clustering

Our study also connects to several key advancements in applying vector representations and topic clustering for analyzing textual data. First, it relates to developing distributed representations of paragraphs, most notably the Paragraph Vector (doc2vec) model introduced by Le and Mikolov (2014). doc2vec extends the word vector framework to sentences, paragraphs, and entire documents by capturing semantic meanings in fixed-length vectors through neural network training. This technique allows for a richer, more context-aware text representation, outperforming traditional bag-of-words models. However, its performance is highly sensitive to various hyperparameters and data configurations, as shown by Lau and Baldwin (2016). Their empirical evaluation of doc2vec provides critical insights into opti-

mizing the model for effective real-world use, emphasizing the importance of careful tuning when applying this method to different datasets.

Adosoglou, Lombardo, and Pardalos (2021) use doc2vec to analyze a vast corpus of 10-K filings from 1998 to 2018. They compare doc2vec with traditional dictionary-based approaches and found that vector representations captured subtle semantic shifts in financial disclosures that were predictive of future abnormal returns. Specifically, they develop the Semantic Similarity Portfolio (SSP) strategy, which identifies firms with minimal year-on-year semantic changes in their filings. Specifically, those with high cosine similarity scores (> 0.95) exhibit significantly higher risk-adjusted returns, up to 10% annually. While their approach showcased the potential of doc2vec in financial text analysis, it also underscored limitations, such as the computational cost of training models and the need to account for executive turnover that might influence document language.

Regarding topic clustering, recent studies have extended the utility of vector-based representations. Calomiris and Mamaysky (2019) apply unsupervised clustering techniques, such as the Louvain method, to group documents by topic and assess their relevance in predicting market outcomes. They create a network of document similarities by constructing vectors of word occurrences for each document and calculating similarity scores. The Louvain method was then employed to detect sub-networks or clusters, which they defined as topics. This clustering approach allowed them to identify significant thematic structures in large corpora of news articles. Such clustered topics could provide valuable insights into future market behavior, particularly in capturing shifts in collective sentiment.

Further reinforcing the benefits of vector-based methods, Curiskis, Drake, Osborn, and Kennedy (2020) compare various document clustering and topic modeling techniques using social media text data. Their findings underscored the effectiveness of document and word embeddings, particularly doc2vec, for document clustering tasks. By outperforming traditional "tf-idf" based approaches and other topic modeling techniques, doc2vec embeddings, combined with k-means clustering, yielded superior results across different datasets. They observe that doc2vec embeddings performed consistently well regardless of document length, although optimal training epoch requirements varied with document size. These results suggest that doc2vec is a robust and adaptable tool for clustering tasks across various text lengths and types.

2.2. Cyber risk and expected stock returns

Ultimately, our study contributes to a growing body of literature that examines the intersection of cybersecurity risk and asset pricing. Jamilov, Rey, and Tahoun (2023) develop a dictionary-based measure of cyber risk exposure using quarterly earnings calls from over 13,000 firms in 85 countries from 2002 to 2021. Their dictionary of cyber-related terms is validated by demonstrating predictability for future cyberattacks and correlations with stock market outcomes and realized volatility. In particular, the study indicates that cyber risk measures can forecast cyberattacks in subsequent quarters and document geographic patterns in cyber risk exposure. The findings are significant for global investors, as U.S. equity holdings in foreign countries predict the cyber exposure of the destination countries. Additionally, the authors examine the pricing of cyber risk in the options market, demonstrating that firms with higher cyber risk exposure incur higher costs for market-based protection against price and variance risks. These results suggest that cyber risk has real economic implications and is increasingly integrated into stock market expectations.

Florackis et al. (2023) present a cybersecurity risk measure based on textual analysis of the "Item 1. A. Risk Factors" section of 10-K filings from 2007 to 2018. By comparing these risk factors across firms using cosine similarity measures, they effectively identify companies with significant cybersecurity risk exposure. The measure correlates with various firm characteristics, such as size, growth opportunities, and R&D expenditures, and predicts future cyberattacks. Furthermore, the authors show that portfolios comprising firms with high cybersecurity risk scores yield a significant annual return premium of up to 8.3%. This risk premium is robust across various specifications, confirming that investors require compensation for holding stocks that are exposed to cybersecurity risks. The study also demonstrates that the cybersecurity-based portfolio underperforms during periods of heightened investor attention to cybersecurity but delivers high returns during other times, suggesting an element of market underreaction to latent cyber risks.

Taking a different approach, Celeny and Maréchal (2023) introduce a method for estimating cyber risk using doc2vec trained on the MITRE ATT&CK cybersecurity knowledge base and applied it to 10-K statements. Unlike the dictionary-based approaches, their model captures broader contextual information from the entire document, resulting in a more comprehensive and accurate cyber risk score. They demonstrate that portfolios sorted by their cyber risk score achieve substantial excess returns—up to 18.72% annually with a significant

risk premium of 6.93% on a long-short portfolio. Their analysis highlights the critical role of cyber risk in the cross-section of stock returns, and their doc2vec-based approach outperforms traditional dictionary-based methods, such as the one developed by Florackis et al. (2023). The study's robustness tests further validate the technique, showing that the cyber risk factor remains consistent over time and is not influenced by the exclusion of cybersecurity firms, thereby providing a more accurate measure of latent cyber risk exposure.

Additionally, Liu, Marsh, and Xiao (2022) explore how firms' sensitivity to cybercrime impacts the pricing of individual stocks and equity portfolios. Using a news-based cybercrime index and corroborating their findings with Google search trends, the authors demonstrate a significant negative correlation between cybercrime exposure and subsequent stock returns. Their study reveals that firms with higher sensitivity to cybercrime tend to underperform in the market, and they further highlight how corporate governance, IT investments, and industry dynamics shape firms' vulnerability to cyber threats. Moreover, they find that high cybercrime beta stocks consistently outperform their low-beta counterparts, particularly following significant cyber incidents. These findings emphasize the importance of incorporating cybercrime exposure into asset pricing models. Jiang, Khanna, Yang, and Zhou (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.

Lastly, Gomes, Mihet, and Risbabh (2023) examine how cyber risk drives firm-level innovation, especially in firms that develop cybersecurity measures internally. Using the cyber risk score from Florackis et al. (2023) alongside patent data, it is evident that firms heavily exposed to cyber risk are incentivized to innovate in cybersecurity solutions, which can contribute to long-term growth.

3. Data and methodology

3.1. Market data

We download public equity data from Wharton Research Data Services[1] (WRDS), and their API. The data originated from the Center for Research in Security Prices[2] (CRSP) and S&P Global Market Intelligence's Compustat database[3]. We report the list of variables in Table A1 and Table 1 report their statistics after cleaning.

We use a pre-existing Python script that retrieves all available data from WRDS about various firms and filters out those that have not filed 10-K forms with the SEC. We extract monthly stock returns and financial ratios for 7,079 firms spanning the period from January 2007 to December 2023. We depict the industry distribution of these firms using the Fama-French 12 industry distribution in Figure 1.

We also download the one-month Treasury bill rate and returns on the market, book-to-market (HML), size (SMB), momentum (UMD), investment (CMA), and operating profitability (RMW) factors from the Kenneth French data repository[4].

[Insert Table 1 here]

[Insert Figure 1 here]

3.2. 10-K statements

10-K statements are financial filings publicly traded companies submit annually to the U.S. Securities and Exchange Commission (SEC). They contain information such as companies' financial statements, risk factors, and executive compensation. 10-K statements will later be used to build a cybersecurity risk measure. The index files from the SEC's Edgar archives[5] are used to download and structure the 10-K. 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 10-K and is structured as follows:

[1] https://wrds-www.wharton.upenn.edu/
[2] https://crsp.org/
[3] https://www.marketplace.spglobal.com/en/datasets/compustat-financials-(8)
[4] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[5] https://www.sec.gov/Archives/edgar/full-index/

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

Where Filename is the URL under which an HTML version of the document is available. Their Central Index Key (CIK) is used to identify firms. The CIK consists of a number used by the SEC to identify corporations and individuals who have filed disclosures. We use a Python script that goes through these index files and identifies URLs corresponding to 10-K statements using the Form Type entry. These URLs are matched to one of the 7,079 firms using the CIK entry. 64,988 10-K statements are identified, corresponding to 2.73 statements per firm on average. Although 10-K filings are annual, portfolios are updated quarterly using the most recently available cyber score derived from the latest 10-K filing. No additional scores are computed from 10-Q filings. The evolution of the number of 10-K filings annually is reported in Figure 2.

[Insert Figure 2 here]

3.3. MITRE ATT&CK description

The MITRE ATT&CK[6] cybersecurity knowledge base is used as a reference for cyberattack descriptions. This knowledge base was established in 2013 to document cyber attack tactics, techniques, and procedures. It is structured by tactics, techniques, and sub-techniques as depicted in Figure 3. 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. Two examples of sub-techniques are given in Table 2.

[Insert Figure 3 here]

[Insert Table 2 here]

Figure 3 and Table 2 are taken from Celeny and Maréchal (2023). The Data section closely follows their approach, and much of their code has been repurposed to suit our requirements. The additional data primarily originates from 2023.

[6]https://attack.mitre.org/

3.4. Cyber score

To compute the cyber scores of interest, we start with the 14 individual MITRE ATT&CK tactics: Reconnaissance, Resource Development, Initial Access, Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Lateral Movement, Collection, Command and Control, Exfiltration, and Impact. It is important to clarify what the cyber score measures. The score is constructed as a semantic similarity measure between firm disclosures and descriptions of cyber attack techniques MITRE ATT&CK. As such, it reflects the extent to which firms discuss concepts related to cyber threats. This includes references to potential vulnerabilities, preventive measures, regulatory disclosures, or past incidents. Therefore, the cyber score should be interpreted as a proxy for cyber risk exposure, rather than a direct measure of realized cyber incidents. To reduce this dimensionality, we aggregate them with clustering methods (see 3.4.4) that yield the following "supertactics": Command and data manipulation, Credential movement, Persistence and evasion, Preparation and reconnaissance. For comparison, we also include the overall score, aggregating all 14 categories into one, corresponding to the score obtained in Celeny and Maréchal (2023). Finally, we add a variation of the overall score that relates better to the risk notion: the cyber sentiment score.

3.4.1. Preprocessing

Everything related to text processing and its use is done exactly as described in Celeny and Maréchal (2023). We download 10-K statements from the SEC Archives as HTML files. Then, we use the library BeautifulSoup to extract usable texts from HTML.[7] We remove punctuation and numbers and set all letters to lowercase. Finally, we apply the Python script of Celeny and Maréchal (2023) that uses the "wordfreq" and NLTK libraries to divide the text into sentences, remove stop-words such as "the", "is", "and", …) and remove the most common words.[8][9]

After pre-processing, the average length of the MITRE ATT&CK sub-technique descriptions is 39.7 words. We utilize a Python algorithm to merge consecutive sentences from 10-K statements into paragraphs with an average length of approximately 40 words after

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

pre-processing. This results in an average of 640 paragraphs per 10-K statement with 46 words per paragraph. The standard deviation is 2.8 words per paragraph and 309 paragraphs per 10-K statement.

3.4.2. Paragraph Vector algorithm (doc2vec)

As in Celeny and Maréchal (2023), we use the paragraph to vector model proposed by Le and Mikolov (2014), which is an extension of the word2vec model (Mikolov, Chen, Corrado, and Dean, 2013). There are various advantages to working with this NLP approach compared to others, such as the dictionary approach. First, the comprehension of the method is semantical, meaning that it is not limited to a count of word frequencies. The word order affects the resulting vector, and paragraphs with similar or synonymous words will have closely related vector representations. Second, training the model with specific text that involves a particular vocabulary allows the incorporation of relatively unknown words. Finally, the resulting vectors have a dimension usually much smaller than the vectors resulting from the dictionary approach.

Two model versions exist: the distributed memory model (DM) and the distributed bag-of-words model (DBOW). In the DM, a neural network is trained as follows. First, a word is removed from a paragraph. Then, by inputting the paragraph vector representation and the context words (also in vector representation) surrounding the missing word, the neural network is optimized to guess the missing word. In the DBOW, the neural network is trained to predict a series of words sampled from a paragraph using only the vector representation of the paragraph as input. Figure 4 illustrates the training process of the two models.

[Insert Figure 4 here]

The training data and details, the hyperparameters and their validation, and the final model choice are extensively covered in Celeny and Maréchal (2023). This work uses their saved doc2vec model.[10]

[10] `https://github.com/technometrics-lab/17-Cyber-risk_and_the_cross-section_of_stock_returns`

3.4.3. Cosine similarity

Using the doc2vec method previously described, all paragraphs of interest can be embedded into vectors. A common way to attribute a similarity score to two paragraph vectors is to take the cosine of the angle they form. Other methods exist, but only measuring the angle has been proven more effective than considering the magnitude of the vectors (see Adosoglou et al., 2021). This is because the latter is more susceptible to the random initialization of weights during training in the neural network that generates the vectors.

3.4.4. Cyber tactics clustering

We disentangle the overall cyber score obtained in Celeny and Maréchal (2023). The idea is that the risk coming from different areas of cybersecurity may not be similarly priced and, therefore, should not be aggregated into a single score; instead, it should be separated into sub-scores of cybersecurity. A natural way to split the overall score into different categories is derived from the written structure of MITRE ATT&CK, with 14 categories already mentioned in chapter 4.1. However, it is believed that splitting the overall score to such an extent might result in a loss of explanatory power and highly correlated sub-cyber scores. Therefore, aggregating the 14 tactics into a few super tactics might mitigate the adverse effect of splitting the overall score.

On the other hand, with the doc2vec method and the similarity score, we can transform every 785 sub-techniques (paragraphs) of MITRE ATT&CK into vectors and compare their similarity. This process yields a similarity matrix of dimension 785 by 785, onto which clustering methods can be applied. Indeed, the similarity matrix can be understood as the representation of a network where every 785 nodes (paragraphs) are connected by edge values weighted by their similarity. In this context, we present three classical clustering methods.

The first and most simplistic clustering method is K-Means. Note that since the input similarity matrix is based on cosine similarity, it is instead designated as spherical K-Means, where the distance between each point and class into K categories is understood as the angle between the vectors defined by those points rather than the Euclidean distance between those points. Either version of K-Means works as follows: It begins by randomly setting initial cluster centroids, then iteratively assigns each data point (paragraph) to the nearest centroid and updates the centroids by recalculating their mean positions among their associated

data points. The process is repeated until convergence. Note that although the K-Means algorithm always converges, it is relatively dependent on the initial centroid guess. The user must choose the number of clusters K without prior knowledge. The algorithm generally produces rough results but often reveals an initial simple structure in the similarity of the provided data.

The second method is much more potent as it requires no prior hyperparameters; thus, the number of clusters is an output of the process. The Louvain method, explained in Blondel, Guillaume, Lambiotte, and Lefebvre (2008), provides a straightforward way to identify clusters (groups of nodes within a graph that are more densely connected) in a network. To explain the Louvain method, we first need to introduce the notion of modularity. It is defined as a value in the range $[-1/2, 1]$ that measures the density of links within communities compared to those between communities. For a weighted graph, the modularity is defined as:

$$Q = \frac{1}{2m} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[S_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where S_{ij} represents the edge weight between nodes i and j, in this case, this is the similarity matrix. k_i and k_j are the sum of the weights of the edges attached to nodes i and j, respectively. m is the sum of all the graph's edge weights. N is the total number of nodes in the graph. c_i and c_j are the communities to which the nodes i and j belong and δ is the Kronecker delta function. The Louvain method works as follows. Initially, each node is assigned to its own community. Then, the process iterates through two phases: the first phase optimizes modularity locally by moving individual nodes between communities to maximize the increase in modularity. The second phase aggregates the nodes in each community from the first phase into a single node and builds a new network, where the communities identified in the first phase are treated as nodes. Phases one and two are repeated until no further improvement in modularity is possible. The final partitioning of nodes into communities is returned as a result.

The third clustering method is spherical K-means on a dimensionally reduced similarity matrix. The spectral clustering method works as follows. First, the degree matrix D is constructed. it consists in a diagonal matrix where each entry D_{ii} represents the sum of similarities for node i and is computed as $D_{ii} = \sum_j S_{ij}$. The Laplacian matrix L is defined

as $L = D - S$. The spectral clustering algorithm computes the eigenvectors and eigenvalues of the Laplacian matrix L. Let $\lambda_1, \lambda_2, ..., \lambda_N$ be the eigenvalues and $v_1, v_2, ..., v_N$ be the corresponding eigenvectors. After obtaining the eigenvectors, we select the K eigenvectors corresponding to the K smallest eigenvalues (excluding the smallest eigenvalue, typically zero). We arrange these eigenvectors as columns in a matrix V of dimension K by N. Finally, we cluster the rows of the matrix V using the k-means clustering method. The power of this approach is that we can choose the number of features necessary to perform a satisfying clustering (reducing from N=785 to K<20, for example, can radically improve the clustering by getting rid of superfluous dimensions).

Finally, scoring is necessary to determine the best clustering output among the wide range of hyperparameters and method choices. We propose a relatively simple but efficient approach that requires initial labeling of each node. Each paragraph (node) is a sub-technique belonging to one of the 14 tactics of MITRE ATT&CK. Thus, they naturally already belong exclusively to 14 sub-clusters. The discrimination of clustering methods works on the following two requirements.

First, we want the paragraphs belonging to one tactic (sub-cluster) to belong to the same super-tactic, *i.e.*, the same cluster found by the method. Indeed, the paragraphs are initially classified by the creator of MITRE ATT&CK together because they share common characteristics. It would not be very sensible to spread them across different super-tactics (clusters) once the clustering method is applied. Thus, a measure of sub-cluster heterogeneity among clusters is needed. We use Shannon entropy, defined as follows:

$$H_{sub_j} = - \sum_{i=1}^{nb.clusters} P(sub_j)_i \log P(sub_j)_i \tag{1}$$

Where $P(sub_j)_i$ is the proportion of paragraphs of sub-cluster (tactic) j belonging to cluster (super tactic) i. Intuitively, if we are in an ideal case and the paragraphs of a sub-cluster j are entirely contained in cluster 1 we would have $P(sub_j)_1 = 1$ and $P(sub_j)_i = 0$ for $i \neq 1$, thus leading to $H_{sub_j} = 0$ being minimal (mind the minus sign in the equation and the logarithm on number lower than 1). If we start to spread the paragraph of the sub-cluster among other clusters, the $P(sub_j)_i$ becomes different from 0 and 1, and H_{sub_j} gradually increases. To reduce the 14 H_{sub_j} scores to one measure of discrimination, we sum them all, thus obtaining the Entropy sum, the sub-cluster heterogeneity measure among clusters. The heterogeneity

is high when the Entropy sum is low.

Second, we need a score to counter the following extreme case. All sub-clusters, except one, can be classified into a single cluster, and the last sub-cluster can be classified into a second cluster. This would lead to a minimum Entropy sum of 0, but would have no value for our application. We want the sub-cluster to be reasonably spread out among the clusters. To translate this idea into a meaningful score, we create a Balanced score, defined as the standard deviation of the label counts. In other words, the clustering method produced an ordered list of 785 values corresponding to the label of the cluster each paragraph belongs to. For each label, we count the number of occurrences on the list. If the paragraphs are relatively well spread out across the cluster, then taking the standard deviation of all the counts of the labels should be low since each cluster would contain approximately the same number of paragraphs. The last case to worry about is that the balanced score could be low, but the paragraph would be randomly spread across the cluster, thus not reflecting the initial structure of MITRE ATT&CK tactics (sub-clusters). To counter that, it is sufficient to consider the Entropy sum.

Considering the method that outputs the lowest Entropy sum and the lowest balanced score, we discriminate the different clustering methods' outputs. Note that there is no guideline regarding the optimal trade-off between the two scores, *i.e.* what additional amount is optimal to forfeit to the Entropy sum to lower the Balanced score and vice versa.

Finally, the entire clustering process described here should be viewed more as a guideline tool. Indeed, after choosing the best method, we classify each paragraph in the cluster where most of its sub-cluster belongs, regardless of the method's output for the misplaced paragraphs. The structure of MITRE ATT&CK is probably more coherent than the output of any unsupervised clustering method. However, it is still advantageous to consider the new clustering structure output, as it is based on the cosine similarity matrix, which could maximize the likelihood of reducing the correlation between the different sub-cyber scores, also based on cosine similarities.

3.4.5. Setting the cyber score

At this point, each paragraph of a 10-K can be transformed into a vector, and the same can be done with the 785 paragraphs of MITRE ATT&CK. Then, each paragraph of the 10-K can be compared to each paragraph of MITRE ATT&CK. This leads to each

paragraph of the 10-K being associated with 785 cosine similarities. Celeny and Maréchal (2023) computes the cyber score of a 10-K by associating the maximum out of the 785 cosine similarities to each paragraph and then taking the average of the top 99% of these maxima. One limitation of this approach is that it does not explicitly distinguish between different contexts in which cyber-related language appears. For instance, mentions of cyber threats may reflect actual incidents, preventive strategies, or general risk disclosures. While this is a common limitation in textual analysis approaches, it may introduce noise in the measurement of cyber risk. Future work could address this by incorporating supervised classification techniques to distinguish between different types of cyber-related statements.

Similarly, we define a sub-cyber score by associating with each paragraph the cosine similarities of a subset of paragraphs of MITRE ATT&CK. For example, each paragraph would be related to 120 cosine similarities (instead of 785), where 120 would correspond to the 120 paragraphs of MITRE ATT&CK that belong to the same category (cluster or sub-cluster/ super tactic or tactic). Then, finding the sub-cyber score associated with a super tactic or tactic would be the same as described in the previous paragraph; we take the maximum out of the 120 cosine similarities for each paragraph and then compute the average of the top 99% of these maxima.

3.4.6. Sentiment analysis

To establish a cyber sentiment score, we opt for a straightforward approach. We define the cyber score as described in Celeny and Maréchal (2023), but instead of taking the maximum, we assign 0 if the paragraph does not contain a word from a specific list and the maximum as usual if it does contain a word from the specific list.

The specific list is defined in Hassan et al. (2019) and is reported in the annex. As shown in the results section, this additional filtering does not improve the empirical performance of the cyber score, suggesting that restricting attention to explicitly risk-related vocabulary may remove relevant contextual information. It contains words related to “risk” or “uncertainty” and was created using the Oxford English Dictionary.

3.5. Asset pricing tests

3.5.1. Univariate sorts

Five portfolios are constructed based on a cyber score of interest. Firms are classified each quarter based on their most recent known cyber score from the previous quarter. These firms are then divided into five categories corresponding to the quintiles of their cyber scores. Consequently, the firms in the top 20% of cyber scores are placed in Portfolio 5 (P5). After that, each firm is weighted within its portfolio according to its market capitalization, known from the end of the previous quarter. The cyber-based portfolios are updated quarterly.

A first quantitative test involves observing whether the average returns of each portfolio change monotonically with increasing cyber score. The idea is to determine if returns are affected by this cyber classification, thereby suggesting a potential cyber-related risk structure.

Next, we assess portfolios' returns, controlling for pricing factors, by using pricing factors recognized in the literature (factors included in the CAPM, in Fama and French, 1992 (FFC) and in Fama and French, 2015 (FF5)), we observe if their linear combinations are sufficient to explain the returns of the portfolio or if statistically significant alpha (intercept) appear, meaning that the profitability of the portfolios based on cyber score can not be entirely explained by common pricing factor and new ones are needed.

3.5.2. Double sorts

The interest in the double sorting method is the same as in univariate sorting. We want to see if returns are affected by the cyber classification. However, the cyber score may be a proxy *i.e.* something that mimics another firm characteristic, such as the size, the book-to-market ratio, or market beta. To avoid that, we sort the firms according to one of the three characteristics mentioned. These firms are then divided into five categories corresponding to the quintiles of their characteristic. Consequently, the firms in the top 20% of the characteristic of interest (for example, firms with the highest book-to-market ratio) are placed in category 5 (Q5). Then, for the firm of each category, Q1 to Q5, we construct a portfolio based on a cyber score as described previously to obtain 25 portfolios, five for each category.

3.5.3. Cross-sectional tests

The empirical strategy proceeds in three steps. First, we estimate factor exposures (betas) for individual assets using time-series regressions. Second, we aggregate these betas at the portfolio level based on portfolio composition. Third, we estimate risk premia using cross-sectional regressions following Fama and MacBeth (1973). This structure allows us to evaluate whether cyber risk is priced in the cross-section of stock returns.

Their method is described as follows. First, estimate betas using time series regressions with 2-year rolling windows (24 months). This corresponds to the following regression for each asset i with $t \in [T-24, T]$:

$$R_{i,t} = \alpha_{i,t} + \sum_k \beta_{i,T}^k F_{k,t} + \epsilon_{i,t}, \quad \forall i \tag{2}$$

Each asset return, R_i, is regressed on non-firm-specific pricing factors, F_k (including standard factors such as market, size, value, momentum, profitability, and investment obtained from the Kenneth French data library), and the cyber score.[11] Consequently, we obtain a time series of betas specific to both asset and factor: $\{\beta_i^k\}_{T=01/2009,\ldots,12/2023}$ (ranging from January 2009 to December 2023, in this example). Then, we build twenty portfolios based on the cyber score, analogously to the five cyber score-based portfolios described earlier. Knowing the weight $x_{i,p,T}$ of each asset inside each portfolio through time, we compute the factor exposures of the portfolios:

$$\beta_{p,T}^k = \sum_{i=1}^{20} x_{i,p,T} \cdot \beta_{i,T}^k \tag{3}$$

After that, the risk premia (gamma) are computed for each time t with $p = 1, \ldots, 20$:

$$R_{p,t} = \gamma_t^0 + \sum_k \gamma_t^k \beta_{p,t-1}^k + \sum_j c_t^j \lambda_{p,t-1}^j + \epsilon_{p,t}^*, \quad \forall t \tag{4}$$

In this specification, the factor exposures estimated in the first step are used to explain portfolio-level returns, while the cyber score enters as an additional firm characteristic aggregated at the portfolio level. Consequently, to determine each $\{\gamma_t^k\}_{t=01/2009,\ldots,12/2023}$, 20

[11] The Kenneth French data library is available at:
https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

portfolio returns are used each time in the linear regression. The additional terms are aggregated firm-specific factors. In our case, we only have one such factor (so j is omitted in the following expression), the cyber score:

$$\lambda_{p,t} = \sum_{i=1}^{20} x_{i,p,t} \cdot \lambda_{i,t} \tag{5}$$

The remaining coefficient $\{c_t\}_{t=01/2009,\ldots,12/2023}$ is determined alongside $\{\gamma_t^k\}_{t=01/2009,\ldots,12/2023}$ during the linear regression. Finally, a t-test is applied on each time series $\{\gamma_t^k, c_t\}_{t=01/2009,\ldots,12/2023}$ to assess the statistical significance of each risk premia.

3.6. Time-series tests

Gibbons et al. (1989) introduces a statistical test to assess portfolio pricing efficiency:

$$R_{i,t} = \alpha_i + \sum_p \beta_{i,p} R_{p,t} + \epsilon_{i,t}, \quad \forall i, \tag{6}$$

where $R_{i,t}$ and $R_{p,t}$ are assets and portfolio returns, respectively. If the portfolios were carefully selected, they could accurately predict the asset returns, thereby reducing the need for alphas (intercepts that contain contributions to their asset returns not accounted for by the explanatory portfolios). GRS provides a statistical test for this null hypothesis: H_0: $\alpha_i = 0 \quad \forall i$. Cochrane (2005) generalizes this idea by including traded factors F_k instead as explanatory variables and portfolio excess returns as the endogenous one:

$$R_{p,t}^e = \alpha_p + \sum_k \beta_{p,k} F_{k,t} + \epsilon_{p,t} \tag{7}$$

In that case, the GRS score that tests jointly the zero alphas follows an F-distribution:

$$\frac{(T-N-K)}{N} \frac{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}}{1 + \hat{\mu}' \hat{\Omega}^{-1} \hat{\mu}} \sim F_{N,T-N-K}, \tag{8}$$

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

must be estimated with the maximum likelihood estimator (biased version). We perform the tests four times on four series of 20 portfolios, each constructed according to the cyber score, size, book-to-market ratio, and market beta of the involved firms.

3.6.1. Bayesian approach

Barillas and Shanken (2018) introduces three methods to test pricing factors. The first method is closely related to the GRS and commonly tests for zero alpha in pricing factors and portfolio returns. The second method tests whether, for a given set of factors, a subset of those factors is sufficient to price portfolio returns. The third method, on which we focus here, enables us to identify which subset of factors, among a large given set of factors, are the most effective pricing factors. It is a "relative" method, meaning that no returns are required for the test. To produce the test, they first introduce the marginal likelihood associated with a given subset of factors:

$$\mathrm{ML} = \mathrm{ML}_U(f|Mkt) \cdot \mathrm{ML}_R(f^*|Mkt, f) \cdot \mathrm{ML}_R(r|Mkt, f, f^*) \tag{9}$$

Where f are the factors of the subset, f^* are the factors excluded from the subset (but in the general set), and Mkt is the market excess returns. Note that the third term can be ignored; it will later be canceled out since it is common to any subset of factors. $\mathrm{ML}_U(Y|X)$ and $\mathrm{ML}_R(Y|X)$ are based on the equations $Y_{t,n} = \alpha_n + X_t\beta_n + \epsilon_{t,n}$ and $Y_{t,n} = X_t\beta_n + \epsilon_{t,n}$. They can be computed as follows:[12]

$$\mathrm{ML}_U(Y|X) = |X'X|^{-\frac{N}{2}} |S|^{-\frac{T-K}{2}} Q \tag{10}$$

$$\mathrm{ML}_R(Y|X) = |X'X|^{-\frac{N}{2}} |S_R|^{-\frac{T-K}{2}} \tag{11}$$

where $|S| = |\epsilon'\epsilon|$ and $|S_R| = |\epsilon'\epsilon|$ are the determinants of the N × N cross-product matrices of associated OLS residuals (R stand for restricted since $\alpha_n = 0$ is imposed on the second linear equation), T is the number of periods, K the number of factors in the regression (number of columns in X), and N the number of endogenous variable on the RHS of the linear equations (number of columns in Y). For example, $\mathrm{ML}_U(f^*|Mkt, f)$ could be associated with

[12] $\overset{(T\times N)}{Y} = \overset{(1\times N)}{\alpha} + \overset{(T\times K)(K\times N)}{X\ \ \beta} + \overset{(T\times N)}{\epsilon}$, note that α has not the correct dimension here, it is to reflect the fact that α is constant across t for a given n.

$[f^*_{1,t}, f^*_{2,t}] = [\alpha_1, \alpha_2] + [Mkt_t, f_{3,t}, f_{4,t}]\beta + [\epsilon^*_{1,t}, \epsilon^*_{2,t}]$ with $N = 2$, $K = 3$ and β a 3×2 matrix and the general set containing four factors $[f_1, f_2, f_3, f_4]$ (two included, two excluded marked by * in this example). The scalar Q is given by:

$$Q = \left(1 + \frac{a}{a+k}\left(\frac{W}{T}\right)\right)^{-\frac{T-K}{2}} \left(1 + \frac{k}{a}\right)^{-\frac{N}{2}} \tag{12}$$

$$a = \frac{1 + \hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}}{T} \tag{13}$$

$$k = \frac{\hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}}{N}(1 - \text{prior}^2) \tag{14}$$

$$W = T\frac{\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}}{1 + \hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}}, \tag{15}$$

where $\hat{\Sigma}$ is the residual covariance matrix, $\hat{\alpha}$ is the vector of alphas, $\hat{\mu}$ is the vector of average X factor, and $\hat{\Omega}$ is the covariance matrix of X factors. Note that both $\hat{\Sigma}$ and $\hat{\Omega}$ must be estimated with the maximum likelihood estimator (biased version). Finally, the prior is an arbitrary number. Barillas and Shanken (2018) use 1.25, 1.5, 2, and 3 in their empirical test. Intuitively, the prior helps to set k, the expected increment to the squared Sharpe ratio $Sh(X)^2 = \hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}$ from the addition of one more factor. Once the relevant marginal likelihoods are computed, the probability p_j associated with a subset of factors M_j being better pricing factors than other subsets is given by:

$$p_j = \frac{\text{ML}_j \times P(M_j)}{\sum_i \text{ML}_i \times P(M_i)}, \tag{16}$$

where ML_j is the marginal likelihood associated with the subset M_j and $P(M_j)$ is the prior probability of the subset M_j. In general, Barillas and Shanken (2018) advise all prior probabilities to be constant and equal since there is no particular reason to favor a specific subset of factors. Note that the third term in Eq. 9 cancels at this last step.

Following the methodology in Barillas and Shanken (2018), p_j can be computed using subsets, including the cyber score as a factor and others without it, to compare its pricing ability.

4. Results

4.1. Clustering of MITRE ATT&CK

We apply clustering methods to the cosine similarity matrix created from MITRE ATT&CK paragraphs vector embeddings. This allows for identifying the relevant sub-cyber score tied to previously mentioned super tactics (command and data manipulation, credential movement, persistence and evasion, and preparation and reconnaissance). We report the results of various attempts with different clustering methods in Figures 5, 6, and 7. The K-means method provides a coherent but rough initial structure, which we report in Figure 5. Indeed, the paragraphs tend to be well spread across the super tactics (clusters), but at the cost of heterogeneity, with the exclusivity of a tactic in a super tactic being nonexistent. This results in a low balanced score at the cost of entropy, as depicted in Figure 8.

Figure 6 shows the performance of the Louvain method. This method dramatically improves heterogeneity, especially with tactics 5 and 12 (resource development and reconnaissance), which are exclusive to cluster 1 (the super tactic: preparation and reconnaissance). However, not putting a threshold on the imputed similarity matrix component induces the Louvain method to create two superfluous clusters. Hence, we include those restrictions. Indeed, when comparing two paragraphs of MITRE ATT&CK, it is not uncommon to encounter sentences with similar structures for different semantic content. Thus, we reduce the similarity of a highly identical paragraph by increasing the threshold. Conversely, we define a lower threshold such that similarities that are too low, and therefore most likely to be noise reflecting no similarity, are set to zero.

The last method can be seen as a safeguard for the output of the Louvain method. Applying the spectral clustering method, we retrieve the structure previously encountered with higher heterogeneity than with K-means. If the hyperparameters are correctly tuned, the output is similar to that of the Louvain method, particularly for $n = 4$ and $egn = 6$. We report the results in Figure 7. Including more dimensions (higher egn value) adds noise and decreases the clustering quality.

Finally, we select the output of the Louvain method as a baseline to group the tactics without splitting them across super tactics. Although Figure 8 shows that outputs of other methods may be slightly better, we favor the Louvain method since no additional hyperparameter tuning is required.

[Insert Figures 5, 6, 7, and 8 here]

This yields the following cluster/super tactics. We name each of them after their content:

Preparation and Reconnaissance: This super tactic encompasses adversaries' tactics to prepare and gather information before launching an attack. **Impact** involves actions that disrupt, destroy, or manipulate systems and data to achieve the attacker's objectives. **Initial Access** includes techniques adversaries use to gain an initial foothold within a network, such as exploiting vulnerabilities or spear phishing. **Resource Development** entails the acquisition of resources like infrastructure, tools, and credentials necessary to support operations. **Reconnaissance** involves gathering information about the target environment to identify potential entry points and vulnerabilities. **Discovery** refers to techniques used to explore and map the target environment, such as network scanning and enumeration.

Persistence and Evasion: Once inside a target network, adversaries employ these tactics to maintain their foothold and avoid detection. **Persistence** ensures the attacker can maintain access even if the system is rebooted or credentials are changed by installing malware or creating rogue accounts. **Privilege Escalation** involves gaining higher-level permissions to access more sensitive information and critical systems. **Execution** refers to running malicious code on a victim system, often necessary to carry out the attacker's objectives. **Defense Evasion** includes a variety of methods to avoid detection and thwart defensive measures, such as disabling security software, obfuscating code, or using fileless malware.

Credential Movement: This group focuses on techniques used to steal and use credentials to move within a network. **Credential Access** involves obtaining account names, passwords, and other secrets that allow attackers to authenticate themselves as legitimate users. Techniques include keylogging, credential dumping, and brute force attacks. **Lateral Movement** is moving through a network to find and access additional targets or more valuable data. This can be done using remote services, exploiting trust relationships, or leveraging legitimate credentials to access other systems and resources.

Command and Data Manipulation: In this phase, adversaries exert control over compromised systems and manipulate data to achieve their goals. **Command and Control**

involves establishing a communication channel with the compromised environment to issue commands and control malware. This can be achieved through web traffic, DNS, or custom communication protocols with command servers. **Collection** refers to gathering sensitive information from compromised systems, such as capturing screenshots, logging keystrokes, or accessing stored files. **Exfiltration** involves transferring the collected data from the target network to an external location controlled by the adversary, often using encrypted channels or covert methods to avoid detection.

4.2. Cyber scores statistical descriptions

From the identified super tactics, we construct cyber scores using the 10-Ks of each firm over the years. Table 3 presents the statistics related to each cyber score (the 14 tactics of MITRE ATT&CK, the four super tactics, the overall score, and the cyber sentiment score). Although their distribution appears similar, several facts must be considered. First, the statistics are for the whole sample, but the distribution is time-varying as Figures 9, 10, 11, and 12 suggest. This means that cyber scores evolving at different rates could be misrepresented. Second, the cosine similarity implies, in theory, a distribution ranging from -1 to 1, whereas the scores are empirically much more narrowly distributed. Thus, the slight variation observed in Table 3 is more meaningful than simple noise.

Two additional aspects must also be reported. First, some tactics lose relevance in the 10-Ks over time, and evidence of the cyber scores reflecting cyber risk has yet to be presented. However, this first feature is encouraging since the cyber scores are evolving differently, showcasing a shift in cyber-related information in the 10-Ks. Second, the cyber sentiment score has a higher 99th percentile than the other score, implying that taking out non-risk-related scores effectively removes points previously belonging to the top one percentile.

[Insert Table 3 here]

[Insert Figures 9, 10, 11, and 12 here]

We present the correlation between cyber scores at the firm level (non-aggregated) in Figure 13. Unsurprisingly, the correlations between all scores are high, except for the sentiment score, which differs in its construction. This is expected, as all scores originate from the same doc2vec neural network output.

[Insert Figure 13 here]

4.3. Cyber scores and probability of cyber-incidents

Table 4 presents logistic regression results exploring the predictive power of the cyber score on the occurrence of the term "cyber" in 8-K filings within 12 months following the release of a 10-K report. Beyond the aggregate "Overall" score, the table examines six dimensions of cyber risk: Sentiment, Preparation, Persistence, Credential, and Command. The dependent variable remains binary, taking the value of 1 if the term "cyber" appears in subsequent 8-K filings and 0 otherwise.

Despite a relatively low explanatory power (Pseudo R^2 ranging from 0.001 to 0.021 across models), the results indicate that the cyber score is a statistically significant predictor of cyber-related disclosures in all dimensions. These results provide important validation for the cyber score. Although the score is derived from textual disclosures, it significantly predicts future cyber-related disclosures, suggesting that it captures meaningful information about firms' underlying cyber risk exposure rather than merely reflecting noise or generic risk language. Specifically, the overall cyber score coefficient of 11.65 is statistically significant at the 1% level, indicating that higher cyber scores are associated with an increased likelihood of 8-K cyber disclosures. This effect holds consistently across dimensions, with coefficients ranging from 9.25 (Sentiment) to 13.49 (Preparation), highlighting that Preparation-related signals appear to have the strongest association with subsequent cyber incidents.

To contextualize these results, consider increasing the overall cyber score from 0.5 to 0.6. This change translates into an approximate 7.3% increase in the probability of a cyber-related disclosure in 8-K filings within 12 months. The dimensional breakdown suggests that firms scoring higher in Preparation and Credential factors exhibit a greater likelihood of reporting cyber incidents, potentially reflecting organizational readiness and response to identified risks.

These findings align with prior literature emphasizing the role of firm-specific factors in predicting cyber incidents (Gordon, Loeb, Zhou, and Wilford, 2024). Furthermore, the newly enforced SEC regulation (effective January 1, 2024) requiring mandatory reporting of cyber incidents through 8-K filings will provide an opportunity to further validate the robustness of the cyber score and its dimensional breakdown in future analyses.

[Insert Table 4 here]

4.4. Cyber scores and financial characteristics

To ensure that the cyber scores are innovative and not the combination of other existent characteristics of the firm, we present the linear regression of the cyber scores of interest in Tables 5, 6, 7, 8, 9, and 10. Compared to Celeny and Maréchal (2023), we add the following variables: readability, secret, risk length table, volume per market capitalization, and humans per market capitalization. we describe all variables in Table A1. The first three variables were part of the tested explanatory variables in Florackis et al. (2023). Including the risk length table shows the critical improvement made with the cyber score of this paper. Indeed, Florackis et al. (2023) report t-statistics of 40.80 and 20.59 for models 1 and 2, respectively. Those t-statistics are significantly lower, thereby improving the independence of our score from non-semantic variables.

[Insert Table 5, 6, 7, 8, 9, and 10 here]

We include a new explanatory variable in the models with the following underlying idea: If a firm has a limited number of employees with highly valuable assets, those assets are more likely to be technological and could be associated with cyber-risk. This risk would then be reported in the 10-Ks and thus be reflected in the cyber score. This choice proves relevant as the Tables report t-statistics close to 10 for all cyber scores. The negative sign of the coefficient supports the view that the lower the human capital ratio, the higher it should be reflected in the cyber score.[13]

The statistical significance of the other coefficients does not deviate significantly from previous studies, with most variables being statistically non-significant or having the same sign as in Celeny and Maréchal (2023), especially at the firm level. Note that, despite adding new variables with higher t-statistics, the R^2 within is still low. It indicates that additional variables cannot fully account for the differences in cyber scores. Furthermore, different t-statistics are obtained for each cyber score for the same coefficient. This suggests that each score could proxy for an intrinsically different risk for the firm.

[13] Additionally, we compute the covariance and correlation of each cyber score with the idiosyncratic volatility of firms. The results are presented in Appendix A2. We thank Prof. Julien Hugonnier for this comment.

Figure 14 displays the correlation of all cyber scores with the mentioned variables. As expected, variables with generally higher t-statistics tend to correlate more with the cyber scores. However, the correlation with "secret" must be taken cautiously since it is a dummy variable and the cyber score is close to 0.5; the correlation may be spurious or, at best, not informative.

Figure 15 shows the average cyber scores across industries. The overall score is always higher than other scores when controlling for industry. This was already the case in Table 3. The cyber sentiment score displays many fewer differences across industries when compared to other scores. This suggests that the score does not contain additional information or even destroy some of it at the industry level. As mentioned in Celeny and Maréchal (2023), industries that rely more heavily on technology, like Business Equipment or Telephone and Television Transmission, potentially report their cyber risk and thus have higher cyber scores. One can also observe that the different cyber scores vary across the industry, further highlighting the potential changes in the source of cyber risk disclosed in the 10-Ks.

[Insert Figures 14 and 15 here]

4.5. Univariate sorts

We now address the following questions: Does a portfolio sorted according to a given cyber score display a structure in its returns? Are the pricing factors usually found in the literature enough to fully explain (not generate alphas on) the return of the cyber-based portfolios?

Tables 11, 13, 14, 15, and 16 all display increasing average excess returns, along with the associated cyber score, with P5 being the portfolio of firms with the highest cyber score. While average excess returns increase across portfolios, the differences between adjacent portfolios are not always statistically significant. This likely reflects limited statistical power rather than the absence of a monotonic relationship. If a given cyber score is an adequate proxy for the associated cyber risk, it implies that taking additional cyber risk grants additional returns. This idea is further explored in the next section.

In Table 12, we report the average returns of portfolios sorted on the cyber sentiment score. Despite all returns being statistically significant at the 1% level, we cannot observe a

monotonic increase in average returns across the scores, and P3 displays the highest average excess returns.

This could indicate that the score does not reflect any risk or that investors are unaware of the type of risk it reflects. We provide additional insight in the following section.

Second, Tables 11, 13, 14, 15, and 16, show that the linear regression using the pricing factors respectively contained in the CAPM, FFC, and FF5 models all grant a statistically significant alpha for P5 only (with statistical significance at the 5% level for CAPM and 1% level otherwise). Alphas are increasing monotonically across the portfolios after controlling for other sources of risk associated with the pricing factors involved. With this evidence in mind, the alphas could partly reflect the cyber risk of each portfolio. This is, however, not the case for the cyber sentiment score in Table 12, as there is no overall strong statistical significance or increasing trend when portfolios are sorted across this variable.

[Insert Tables 11, 12, 13, 14, 15, and 16 here]

4.6. Double sorts

We aim to determine if organizing quarterly portfolios first based on specific characteristics and then based on the cyber score results in a structure in their average excess returns. The idea is that controlling for additional characteristics rejects the hypothesis that the cyber scores are a proxy for another firm variable. Thus, the cyber score would capture the cyber risk exposure and the associated additional returns. In other words, the increasing cyber score, which should reflect the rising cyber risk, still displays increasing returns related to cyber risk even if the set of firms to analyze is already organized and structured according to another characteristic unrelated to cyber risk.

Table 17 displays the average returns of various double-sorted portfolios. Notably, there is a clear increasing trend in average returns as the overall cyber score quintile increases, which contrasts with the findings of Celeny and Maréchal (2023), where the trend was less pronounced. In this analysis, only the first quintile of the book-to-market ratio consistently shows no increase in value. Note that the market beta at Q3 and the size at Q5 differ by 0.01%. However, this quantitatively marginal result may be spurious.

For the cyber sentiment score, as previously observed in Table 12, there are no additional returns for an increase in the score, so the cyber sentiment score certainly does not reflect

any risk.

The returns obtained with other cyber scores display an interesting aspect. Command and data manipulation, credential movement, and persistence and evasion strongly suggest a monotonic increasing trend and therefore, cyber risk premia, except for the lower quintile, where the conclusion might seem slightly less evident. However, this is not the case for preparation and reconnaissance, for which the trend is nonmonotonic almost everywhere. This could suggest two things. It is possible that investors do not fully appreciate the risk this cyber score reflects. Alternatively, preparation and reconnaissance reveal no risk.

[Insert Table 17 here]

4.7. Cross-sectional tests

We test whether a cyber score increase drives a return increase in cyber-based portfolios, controlling for other well-known pricing factors using the regression method described in Fama and MacBeth (1973). Table 18, 19, 20, 21, 22, and 23 display the results for each cyber score using a different pricing model that includes the cyber score.

The cyber sentiment score probably reflects no meaningful reality regarding the firms; therefore, constructing portfolios based on it reveals no particular structure, as observed in Table 19. No risk premia are observed for any of the involved pricing factors (including the cyber sentiment score), and no statistically significant alpha exists.

The overall cyber score on Table 18 displays positive additional returns for an increased cyber score that is statistically significant at the 10% level when included as the only explanatory variable or with the market factor and at the 5% level when included with the pricing factors from Fama and French (1992). However, a collinearity problem arises for the fifth model, which incorporates additional factors from Fama and French (2015). The cyber score aggregated for all firms in the cyber-based portfolios constructs a factor that may be collinear with CMA. Therefore, the statistical significance of the cyber score is significantly reduced. When compared to Celeny and Maréchal (2023), they appear not to suffer from collinearity and have a lower adjusted R^2.

The command and data manipulation score in Table 20 yields similar results, also plagued by collinearity. On the contrary, the remaining scores in Tables 21, 22, and 23 do not display collinearity, and their respective cyber score appear statistically significant at the 5% level.

Note that, on all Tables (except for the cyber sentiment score), the coefficients of the cyber score appear positive, further indicating that cyber scores effectively reflect cyber risks that are rewarded on the market. A standard deviation of the overall cyber score (0.03 in Table 3) generates an additional return of $0.03 \cdot 0.04 = 0.12\%$ compared to Celeny and Maréchal (2023) with 0.18%.

[Insert Tables 18, 19, 20, 21, 22, and 23 here]

4.8. Time series tests

We use the GRS test to examine whether adding the long-short portfolio $P5 - P1$ as a pricing factor enhances the explanatory power of the five-factor model of Fama and French (2015). We aim to determine if we can globally reduce the unexplained portion of returns to nearly zero. We conduct four GRS tests on 20 portfolios, sorted quarterly, based on firms' cyber scores, size, market beta, and book-to-market ratios, respectively.

Tables 24, 25, 26, 27, 28, and 29 display the four GRS tests for each cyber score of interest. All tables give similar results, with the probability of alphas being commonly zero increasing as we add the cyber factor P5-P1, and the average R^2 increasing. There is an exception when portfolios are sorted by size. Their associated probabilities appear to decrease, but it is essential to note that the probability was already high before the addition of the cyber factor. It is hard to quantify whether the alphas are closer to zero when they are already commonly near zero. Also, note that the difference in probabilities in the case of market beta is positive but small compared to the improvement provided by the cyber factor when explaining the returns of portfolios sorted on the cyber score or the book-to-market ratio.

[Insert Tables 24, 25, 26, 27, 28, and 29 here]

4.9. Bayesian asset pricing tests

We conduct an additional test to evaluate the cyber factor P5-P1 as a reliable pricing factor. Using the Bayesian GRS (BGRS) test described earlier, we can identify the optimal subset of pricing factors from a large set of potential factors. Figures 16, 17, 18, 19, 20, and 21 present the BGRS test for each cyber score of interest. Keeping only the subsets with the higher probabilities at the end of the time range, one can notice that the top five

subsets always include the cyber factor (except for the cyber sentiment score case, where one of the top five subsets does not include the cyber factor). When considering the cumulative probabilities associated with each pricing factor, it becomes clear that the cyber factors have consistently exhibited a growing trend over the years, indicating that they have become increasingly relevant as a pricing factor over time.

[Insert Figures 16, 17, 18, 19, 20, 21 here]

4.10. Additional tests

In this section, we conduct further tests to expand the range and depth of understanding of the cyber scores we developed. These additional tests aim to identify potential limitations and verify the scores' behavior in a real case.

In Table 30, we display the probabilities associated with the differences in mean returns between the overall and other cyber-based portfolios. Although we give evidence that the cyber scores reflect different realities related to the cyber subject in the 10-Ks, when portfolios are constructed from these scores, their returns do not display any statistically significant variation across different scores. In the context of market perception of risk, the results suggest that, although cyber risk can be decomposed into multiple dimensions, financial markets do not differentiate between them and instead price a single aggregate cyber risk factor.

[Insert Table 30 here]

In December 2020, a significant cyber attack on SolarWinds, a major IT management company, was uncovered, marking one of the most extensive and sophisticated cyber espionage operations. The attackers, believed to be state-sponsored, infiltrated SolarWinds' Orion software, used by numerous high-profile clients, including Fortune 500 companies and various U.S. government agencies. The attackers gained unprecedented access to sensitive data across multiple networks by embedding malicious code in a routine software update. This breach highlighted vulnerabilities within supply chain security and underscored the broader implications for firms at risk of cyber attacks. The incident serves as a case study for analyzing the financial impact on companies deemed to be cyber-risky. Such analysis

using this event was also performed in Florackis et al. (2023), setting December 14, 2020, the day the attack was disclosed to the SEC, as the event day.

We conduct a similar analysis on Table 31. None of the abnormal returns were statistically significant. Figure 22 illustrates the cumulative returns of the cyber-based portfolio around this event. The results are unconventional. Returns were higher in the lower-tier cyber-based portfolio in the days leading up to the event. Moreover, when the event occurred, all portfolios declined except for P5. However, two critical factors need to be considered. First, each portfolio aggregates over 600 firms. The SolarWinds breach may still be too financially localized to impact a large number of firms, and its effect could be diluted among unaffected firms (those not associated with SolarWinds or not perceived by the market as being affected). Second, none of the variations are statistically significant, and opposing behaviors likely mitigate the event's overall impact. For instance, during the shock, investors might have shifted their investments to other stocks considered safe but still related to cybersecurity, or the event might have increased interest in cybersecurity and boosted investment in P5 firms.

[Insert Table 31 here]

[Insert Figure 22 here]

Finally, we display in Tables 32 and 33 the cumulative abnormal returns of P20 and P5 but constructed with different cyber scores. There is not enough statistical significance to infer anything. Note that the cumulative returns still reach higher returns in the P20 case (Figure 23) than in the P5 case (Figure 24), and the portfolios based on the various cyber scores seem to behave similarly, which supports the hypothesis that the market perceives a single aggregated risk related to cybersecurity.

These results contrast those of Florackis et al. (2023), who find a statistically significant drop in their top cyber-based portfolio returns around the event.

[Insert Tables 32 and 33 here]

[Insert Figures 23 and 24 here]

5. Conclusion

In this study, we utilize a doc2vec model to transform paragraphs from the MITRE ATT&CK database's descriptions of cyberattacks into vectors. Comparing those vectors based on their cosine similarity, we apply clustering methods such as K-means, Louvain, and spectral clustering to infer groups of cyber attacks belonging to four defined types (super-tactics): command and data manipulation, credential movement, persistence and evasion, and preparation and reconnaissance. Those clusters were recurrent across different trials using the three methods and different hyperparameters. They were also chosen to preserve the underlying written structures of MITRE ATT&CK by using a two-score system that ensures the equal distribution of paragraphs across super-tactics and their exclusivity to these super-tactics.

Then, we use the doc2vec model to transform paragraphs of annual statements, more precisely 10-Ks, into vectors. Building the cosine similarity between 10-K vectors and vectors belonging to specific super tactics allows me to infer the semantic similarity of the 10-Ks to the four types of cyber attacks. We define those cosine similarities as the cyber score of a 10-K for a given super tactic. We also build an additional cyber sentiment score. This score considers only paragraphs' cyber scores when they contain words related to a "risk" or "uncertainty" vocabulary.

We find that the different cyber scores cannot be explained by the linear combination of standard financial variables and non-semantic variables of the firms they belong to (the highest R^2 within among all tested cyber scores is 0.43). The independence of those newly found variables supports their innovative nature. All aggregate cyber scores have increased over the years and are higher in industry sectors (from the 12 Fama-French industries classification) that involve assisting and workflow-related technologies, such as Telephone and Television Transmission or Business Equipment.

We conduct asset pricing and statistical tests involving portfolios sorted on firms' cyber scores to assess if the cyber scores reflect cyber risks. Since all results for each cyber score are similar to the overall cyber score and the previous study Celeny and Maréchal (2023) only use this aggregated cyber score, we report here only the results related to this score. This does not apply to the cyber sentiment score, for which it appears clear that no risk premium is involved.

Organizing firms into portfolios based on their cyber scores allows for the observation of increasing average excess returns as the portfolio's cyber score increases. The portfolio with the lowest quintile of cyber score, P1, has an average excess return of 0.82%. The portfolio with the highest quintile of cyber score, P5, has an average excess return of 1.44% (both statistically significant at the 1% level). Thus, a long-short portfolio P5-P1 destroys performance. Then, controlling for common pricing factors, we find that P5 has an alpha of 0.29% at the 1% level. Conversely, other portfolios, P1 to P4, have increasing alphas but are statistically insignificant. This threshold in significance between P4 and P5 highlights the fact that we cannot tell from a firm's cyber score at which point it truly highlights cybersecurity in its 10-Ks. Therefore, lower portfolios contain a variety of firms that may be classified according to noise without any meaningful distinction. We recommend that future studies using a similar framework focus solely on P5, rather than P5-P1, as has been done to date. Sorting the firms into a first unrelated category and then according to their cyber score also reveals a similar structure of returns, as previously mentioned, with the top cyber-based portfolios performing better. Thus, the structure is robust, controlling for other firm characteristics.

Fama and MacBeth (1973) regressions show a risk premium associated with the cyber score and all disentangled cyber scores. In contrast, the cyber sentiment score does not drive any risk premium. The GRS test of Gibbons et al. (1989) shows that the long-short portfolio P5-P1 helps to price various assets when used with the other well-known pricing factors of the five-factor model Fama and French (2015). Furthermore, the BGRS tests from Barillas and Shanken (2018) also highlight that P5-P1 is an important cyber-based pricing factor. According to the test, this importance is rising with time. Interestingly, these observations are valid for all cyber scores, including the cyber sentiment score.

Last, we cannot reject the hypothesis that the return of P5-P1, built with different cyber scores, is statistically different. Then, we conduct an event study using the cyber breach of SolarWinds in December 2020. The analysis provided no conclusive results, except that portfolios based on different cyber scores exhibit similar behavior. These last two observations could prove that the market does not differentiate between the various types of cyber risk and perceives them as a single aggregate cyber risk.

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.

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.

Arslan-Ayaydin, O., Boudt, K., Thewissen, J., 2016. Managers set the tone: Equity incentives and the tone of earnings press releases. Journal of Banking and Finance 72, 132–147.

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

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment P10008.

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

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.

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

Celeny, D., Maréchal, L., 2023. Cyber risk and the cross section of stock returns. Available at `http://dx.doi.org/10.2139/ssrn.4587993`

Cochrane, J. H., 2005. The risk and return of venture capital. Journal of Financial Economics 75, 3–52.

Curiskis, S. A., Drake, B., Osborn, T. R., Kennedy, P. J., 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing and Management 57 (2), 102034.

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., 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.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies 15, 915–953.

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.

Gomes, O., Mihet, R., Risbabh, K., 2023. Data risk, firm growth and innovation. Available at: `http://dx.doi.org/10.2139/ssrn.4559921`

Gordon, L. A., Loeb, M. P., Zhou, L., Wilford, A. L., 2024. Empirical evidence on disclosing cyber breaches in an 8-k report: Initial exploratory evidence. Journal of Accounting and Public Policy 46, 107226.

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.

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.

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

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, 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, 1188–1196.

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`

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

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

Tables and Figures

Variable	Mean	Std	Min	Max	P1	P25	P50	P75	P99
firm size	20.17	2.49	13.11	26.33	13.68	18.54	20.33	21.87	25.66
firm age	2.42	1.15	-2.48	4.13	-1.39	1.83	2.67	3.24	4.05
ROA	-0.15	0.54	-4.30	0.49	-3.04	-0.13	0.02	0.06	0.37
book to market	0.73	1.19	0.00	99.55	0.02	0.26	0.50	0.86	4.56
TobinQ	2.11	2.14	0.35	24.15	0.56	1.03	1.39	2.24	11.80
MktBeta	1.15	0.87	-3.00	5.91	-1.15	0.65	1.08	1.55	3.99
intangibles to assets	0.15	0.21	0.00	8.10	0.00	0.00	0.05	0.24	0.78
debt to assets	0.57	0.30	0.03	1.81	0.05	0.34	0.55	0.78	1.48
ROE	-0.08	0.61	-5.88	1.60	-2.96	-0.07	0.07	0.15	0.87
price to earnings	-0.87	132.55	-2001.73	455.35	-568.94	-3.93	12.05	22.64	295.69
profit margin	-0.42	6.46	-111.45	1.00	-27.21	0.22	0.39	0.62	0.96
asset turnover	0.82	0.74	0.00	4.09	0.01	0.25	0.66	1.16	3.42
cash ratio	2.07	3.89	0.01	36.13	0.01	0.23	0.68	1.98	20.28
sales to invested cap	1.39	1.50	0.00	10.42	0.01	0.44	0.94	1.77	8.13
capital ratio	0.31	0.32	-0.10	1.97	0.00	0.03	0.24	0.47	1.51
RD to sales	0.75	5.08	0.00	89.34	0.00	0.00	0.00	0.06	22.67
ROCE	-0.00	0.44	-3.21	1.30	-1.98	-0.01	0.09	0.17	0.93
readability	16.08	1.07	7.14	19.89	13.19	15.51	16.31	16.81	18.08
secret (dummy)	0.28	0.45	0.00	1.00	0.00	0.00	0.00	1.00	1.00
risk length table	5.03	1.46	0.00	7.69	0.00	4.84	5.35	5.81	6.84
volume per cap	0.28	6.52	-1.91	3485.03	-0.01	0.06	0.13	0.23	1.97
humans per capital $\times 10^6$	8.40	157.41	0.00	16689.97	0.00	0.50	1.50	4.19	71.49
humans per assets $\times 10^6$	4.29	14.04	0.00	879.49	0.00	0.38	1.75	4.08	42.02

Table 1: **Descriptive statistics of the firm characteristics**

This table provides descriptive statistics for various firm characteristics from 2009 to 2023. Mean, standard deviation (Std), minimum (Min), and maximum (Max) values are reported. Percentiles (P1, P25, P50, P75, P99) are also included. Firm-level characteristics are winsorized at the 1st and 99th percentile (by year). The characteristics are defined in Table A1

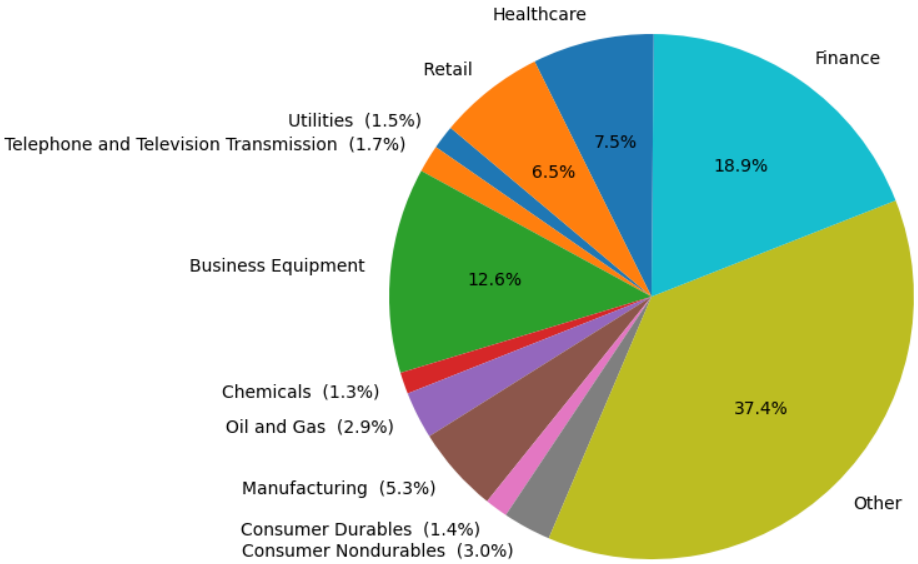

Fig. 1: **Industry distribution**

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

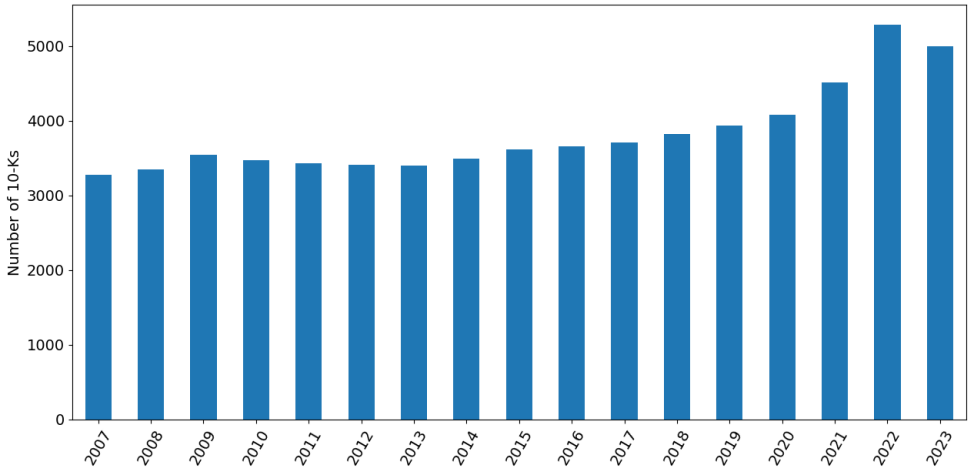

Fig. 2: **Number of 10-Ks per year**

Number of companies in the study sample that have filed a 10-K statement through the years.

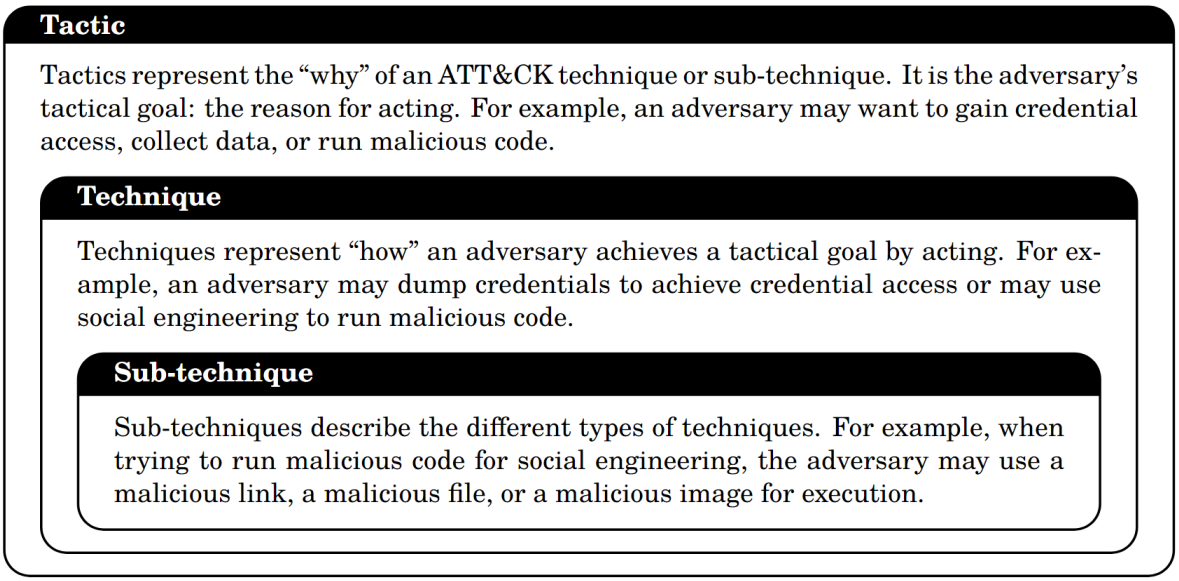

Fig. 3: **Structure of MITRE ATT&CK**

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

Table 2: **Examples of sub-techniques from MITRE ATT&CK**

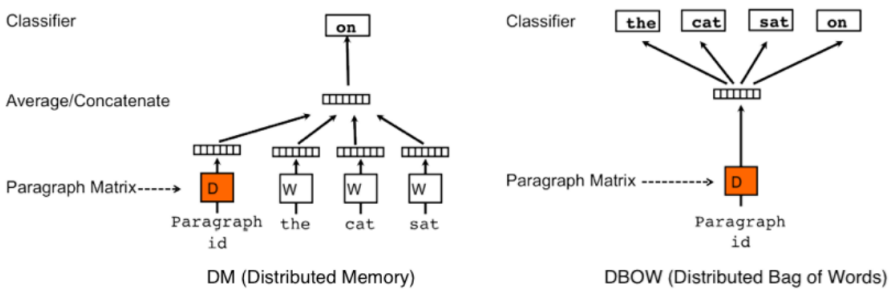

Fig. 4: **Illustration of doc2vec training**

Illustration of the training of the neural network of the two versions of doc2vec, distributed memory model (DM) and distributed bag-of-words model (DBOW). The figure is taken from Le and Mikolov (2014).

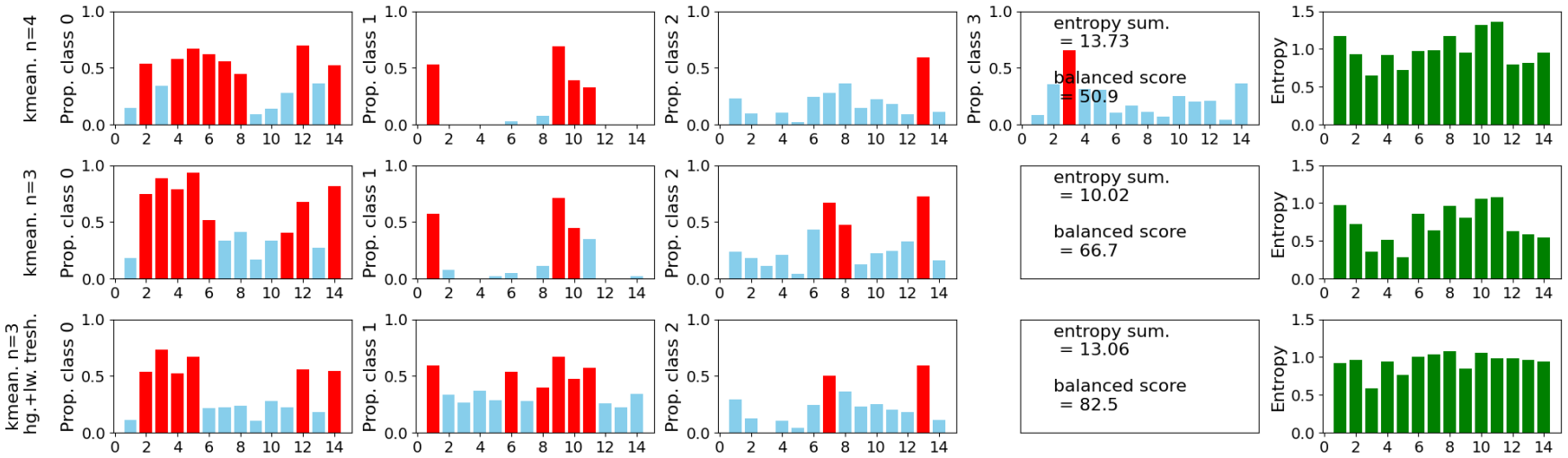

Fig. 5: **Clustering results part.1**

This figure presents the results of each clustering method indicated on the left. The figure in red and blue represents $P(sub_j)_i$, the proportion of paragraphs of sub-cluster (tactic) j belonging to cluster (super tactic) i. The 14 sub-cluster labels are on the x-axis of each figure, and the cluster labels correspond to the columns (class 0 to 3, here). If the proportion is in red, it means it is the highest in the cluster (in other clusters/columns, the same sub-cluster will be in blue). I also report the entropy sum and the balanced score on the figure for each method. Finally, the individual Shannon entropy of each sub-cluster is reported in green in the last column. In the name of the method, I also indicate the hyperparameters of the method. Here, n corresponds to the number of clusters imposed by the k-means method. "hg. tresh." and "lw. tresh." corresponds to a change applied to the similarity matrix. If the value in the similarity matrix is lower than 0.25, it is changed to 0 (lower threshold), and if the similarity is higher than 0.85, it is changed to 0.5 (higher threshold). In part.2 and part.3 "egn" corresponds to the K eigenvectors in the spectral clustering. I also made the output clusters of each method match. Hence, the comparison is simpler (otherwise, what the Louvain method called cluster 2 is not necessarily cluster 2 for the k-means method). The following list shows the corresponding number of each tactic : 1: Persistence, 2: Command and Control, 3: Impact, 4: Initial Access, 5: Resource Development, 6: Collection, 7: Exfiltration, 8: Credential Access, 9: Privilege Escalation, 10: Execution, 11: Defense Evasion, 12: Reconnaissance, 13: Lateral Movement, 14: Discovery.

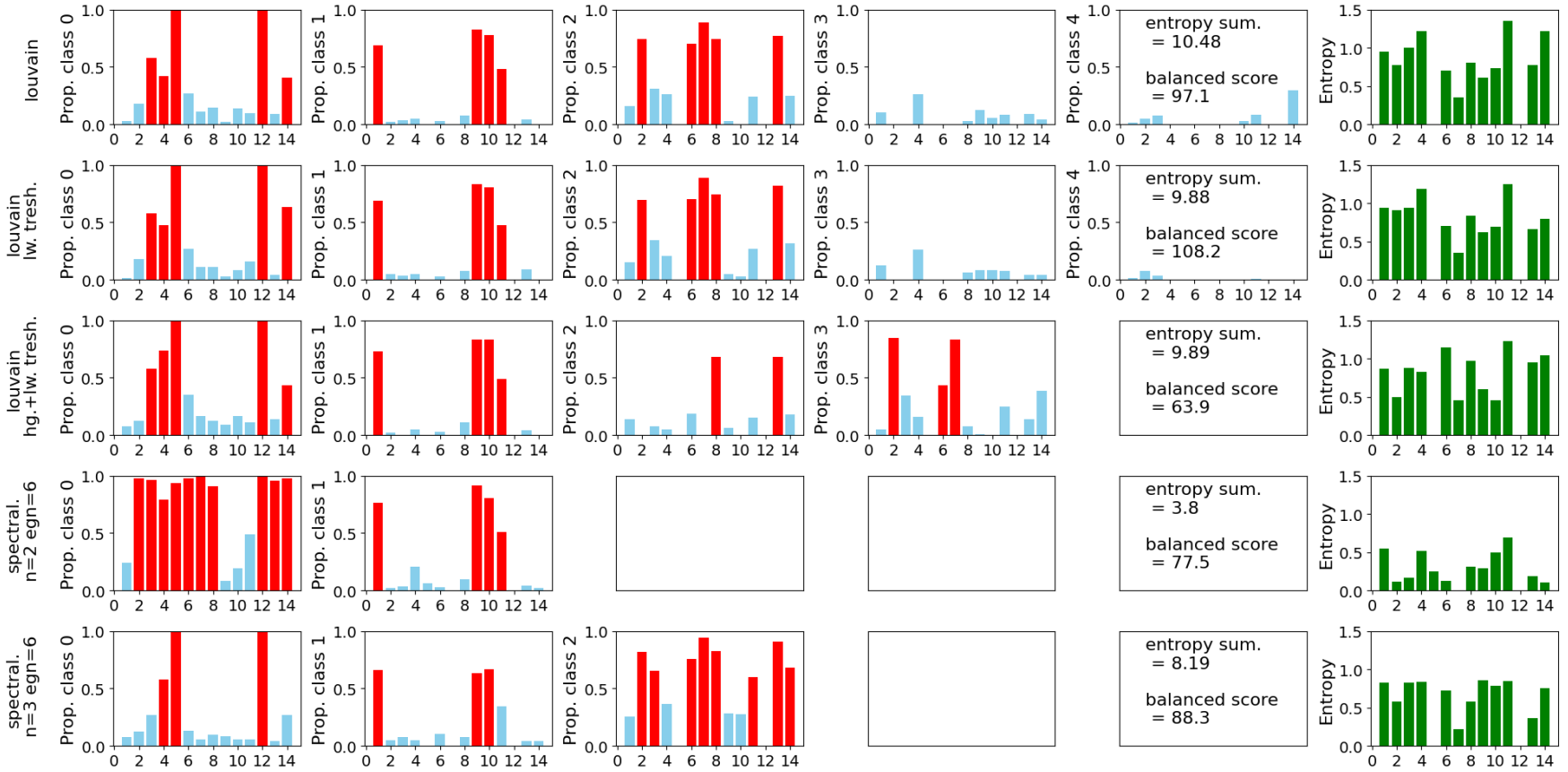

Fig. 6: **Clustering results part.2**

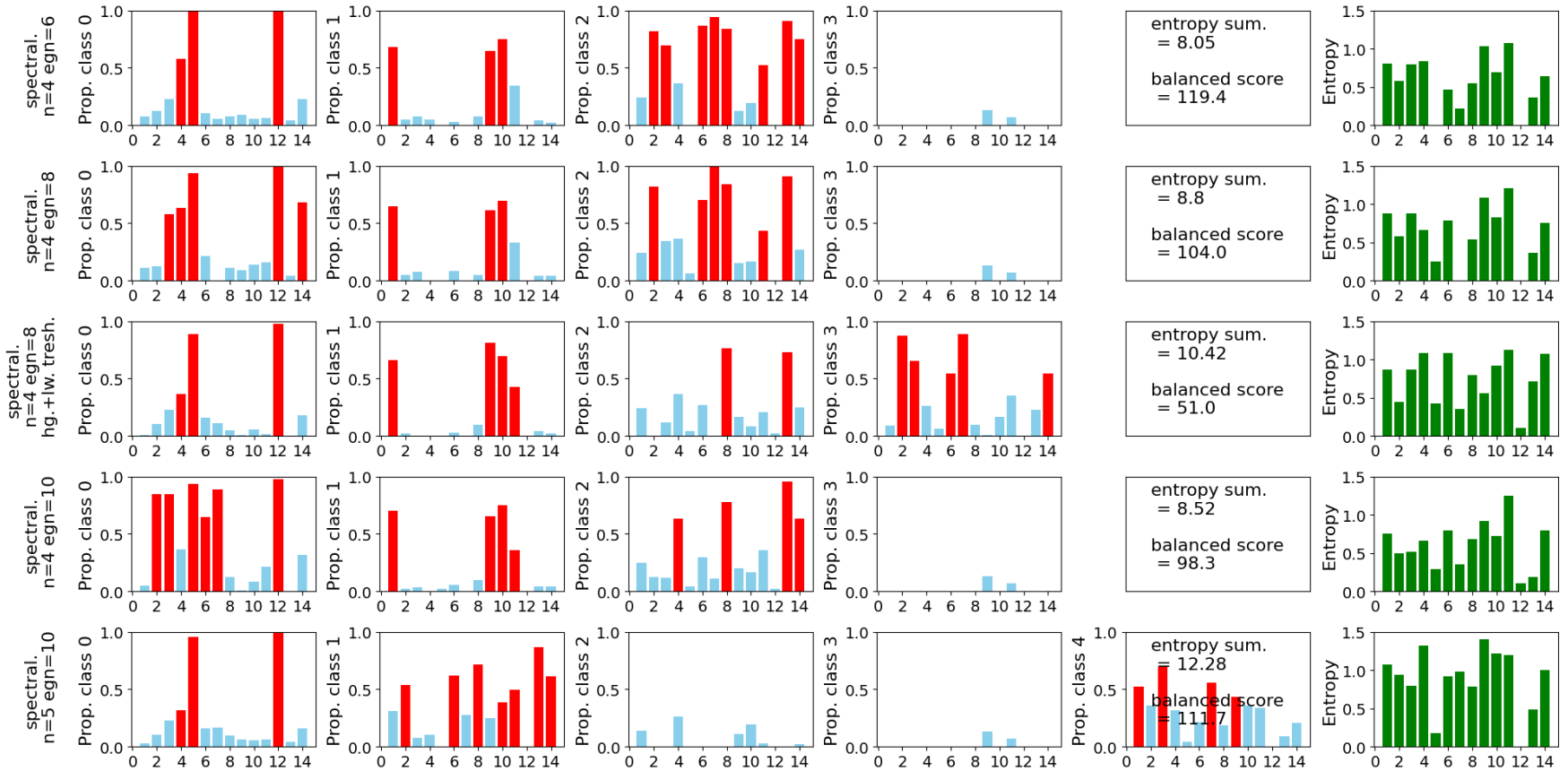

Fig. 7: **Clustering results part.3**

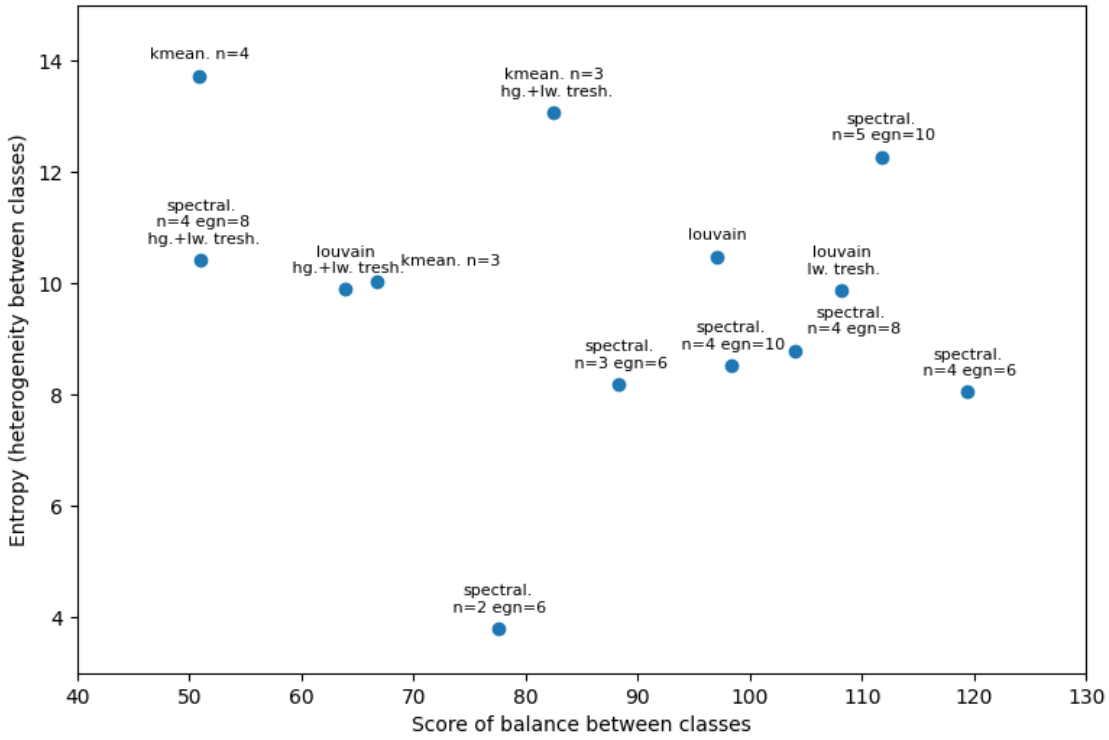

Fig. 8: **Comparison of clustering scores: Entropy sum and Balanced score**

Each clustering method of Figure 5, 6, and 7 is presented here using their respective entropy sum and balanced score. Recall that the aim was to reduce both scores to distinguish the best clustering method. Also, note that there is no guideline regarding what additional amount is optimal to forfeit to the entropy sum to lower the balanced score and inversely.

	Mean	Std	Min	Max	P1	P25	P50	P75	P99
Persistence	0.49	0.03	0.27	0.64	0.44	0.47	0.49	0.51	0.58
Command and Control	0.47	0.03	0.28	0.62	0.42	0.45	0.47	0.49	0.55
Impact	0.47	0.03	0.25	0.59	0.41	0.45	0.47	0.50	0.55
Initial Access	0.46	0.03	0.23	0.59	0.40	0.44	0.46	0.48	0.55
Resource Development	0.47	0.03	0.23	0.62	0.41	0.44	0.47	0.49	0.56
Collection	0.49	0.03	0.29	0.64	0.43	0.46	0.48	0.51	0.57
Exfiltration	0.47	0.03	0.23	0.64	0.41	0.44	0.46	0.49	0.56
Credential Access	0.50	0.03	0.29	0.64	0.43	0.47	0.49	0.52	0.58
Privilege Escalation	0.48	0.03	0.27	0.64	0.43	0.46	0.47	0.49	0.56
Execution	0.46	0.03	0.29	0.61	0.42	0.44	0.46	0.48	0.55
Defense Evasion	0.51	0.03	0.29	0.65	0.46	0.49	0.50	0.52	0.59
Reconnaissance	0.48	0.03	0.32	0.61	0.42	0.46	0.48	0.51	0.57
Lateral Movement	0.47	0.03	0.26	0.64	0.43	0.45	0.47	0.49	0.56
Discovery	0.48	0.03	0.31	0.63	0.43	0.46	0.47	0.49	0.56
Preparation and Reconnaissance	0.50	0.03	0.33	0.64	0.44	0.48	0.50	0.53	0.58
Persistence and Evasion	0.51	0.03	0.29	0.65	0.46	0.49	0.51	0.53	0.59
Credential Movement	0.50	0.03	0.29	0.65	0.44	0.48	0.50	0.52	0.59
Command and Data Manipulation	0.50	0.03	0.29	0.64	0.44	0.47	0.49	0.52	0.58
Overall	0.53	0.03	0.33	0.65	0.47	0.50	0.52	0.54	0.61
Sentiment	0.51	0.05	0.00	0.72	0.42	0.48	0.51	0.54	0.63

Table 3: **Descriptive statistics of cyber scores**

This table provides descriptive statistics for the 14 MITRE ATT&CK tactics cyber score, the four aggregated sub-cyber scores of the super-tactics, the overall cyber score, and the cyber sentiment score. The statistics are computed from all firms from 2009 to 2023.

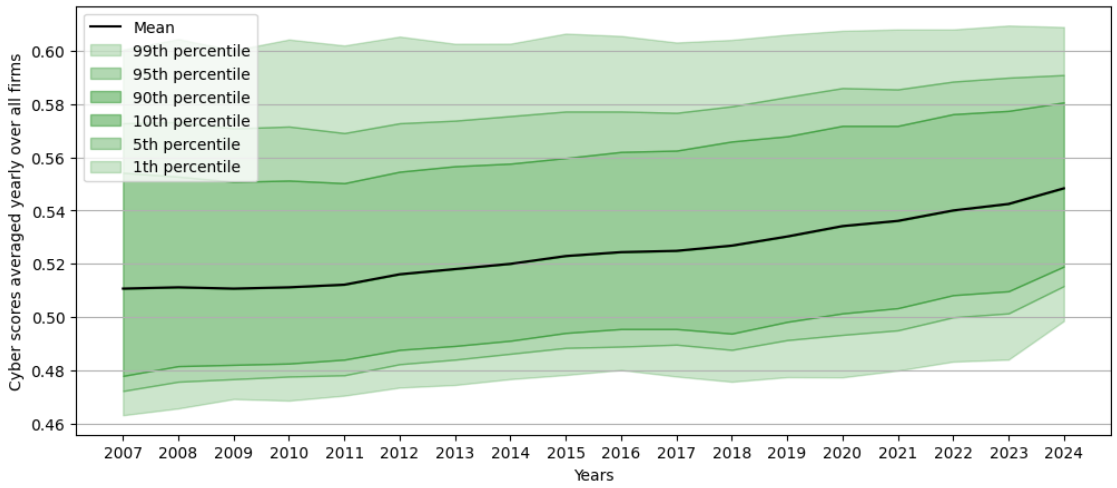

Fig. 9: **Evolution of the overall cyber score averaged yearly over all firms**

The figure shows the evolution of the overall cyber score over all firms yearly. Each year provides a distribution of the cyber score over all firms that can be sorted to provide percentiles of interest and the averaged cyber score for a given year.

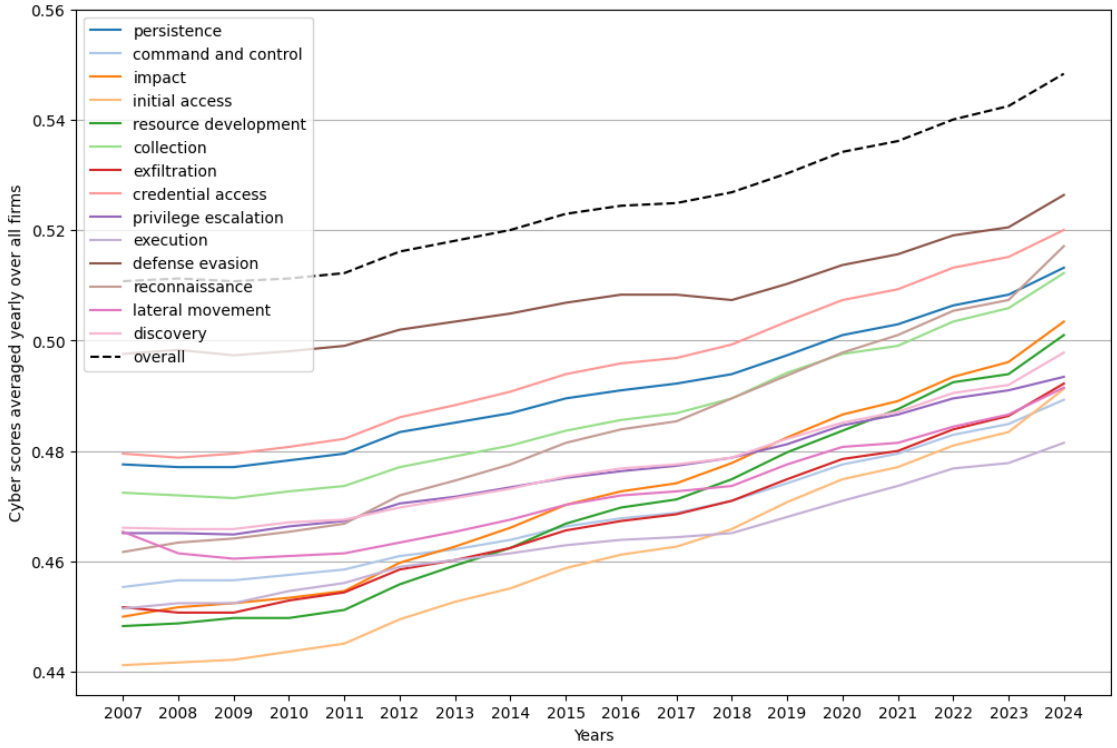

Fig. 10: **Evolution of the sub-cyber scores related to the 14 tactics averaged yearly over all firms**

The figure shows the evolution of the 14 sub-cyber scores averaged over all firms yearly. The overall cyber score is also included to allow comparison.

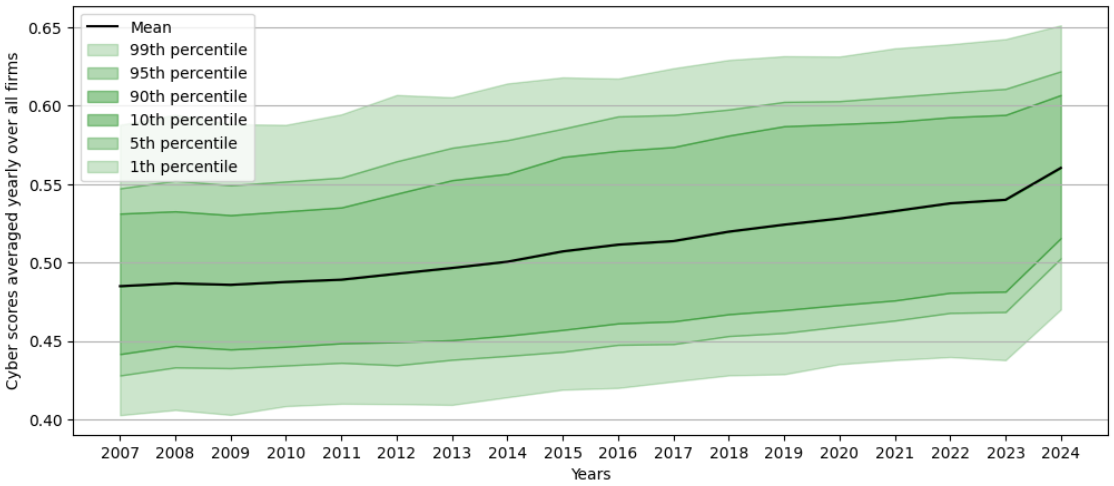

Fig. 11: **Evolution of the cyber sentiment score averaged yearly over all firms**

The figure shows the evolution of the cyber sentiment score over all firms yearly. Each year provides a distribution of the cyber score over all firms that can be sorted to provide percentiles of interest and the averaged cyber score for a given year.

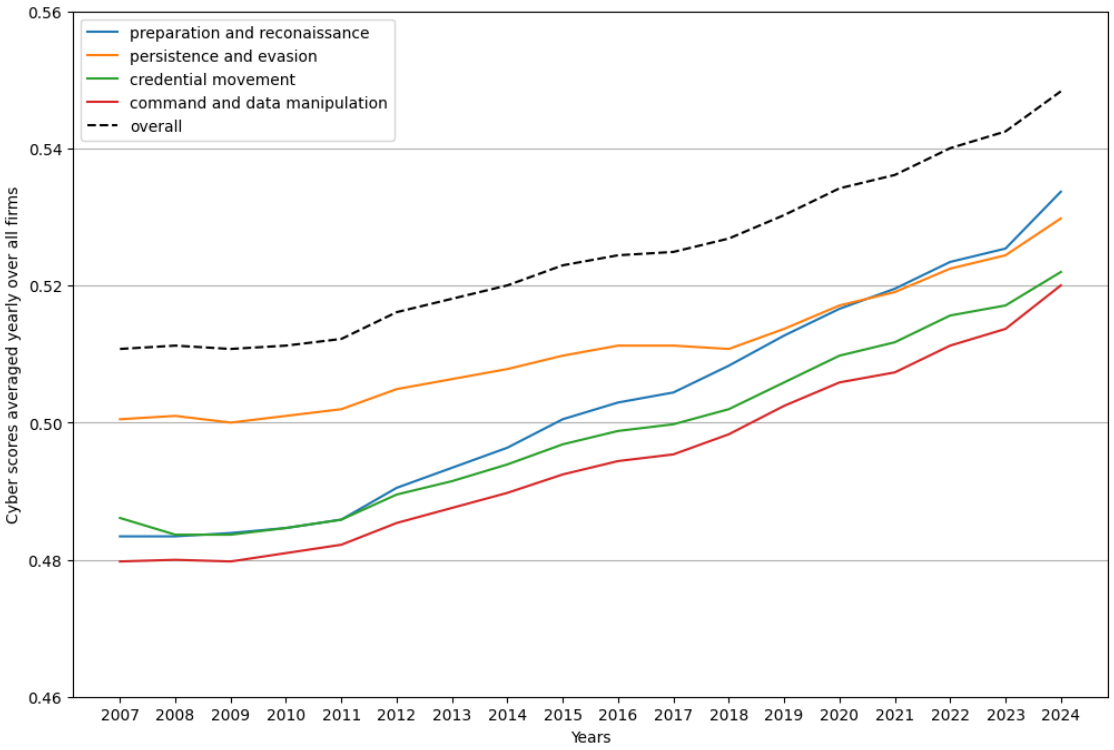

Fig. 12: **Evolution of the sub-cyber scores related to the four super-tactics averaged yearly over all firms**

The figure shows the evolution of the four sub-cyber scores averaged over all firms yearly. The overall cyber score is also included to allow comparison.

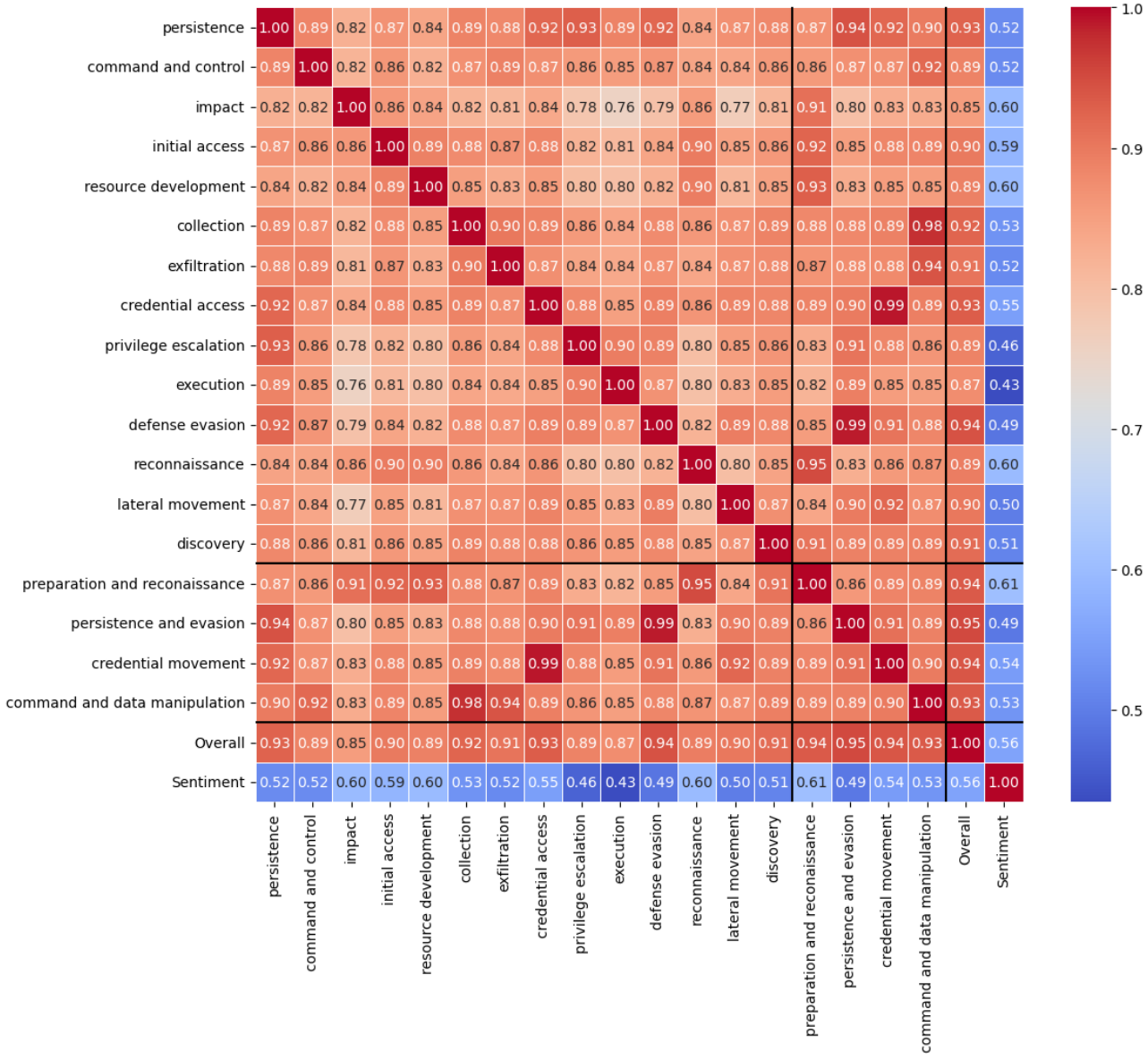

Fig. 13: **Correlations of all cyber scores**

Firm-wise correlations of the sub-cyber scores of the 14 MITRE ATT&CK tactics, the 4 aggregated sub-cyber scores of the super-tactics, as well as the overall cyber score and the cyber sentiment score are presented here.

	Overall	Sentiment	Preparation	Persistence	Credential	Command
Constant	-9.54***	-8.20***	-10.26***	-8.66***	-8.98***	-8.71***
T-statistic	(-23.85)	(-31.16)	(-27.42)	(-21.04)	(-24.96)	(-24.78)
Cyber score	11.65***	9.25***	13.49***	10.28***	11.11***	10.65***
T-statistic	(15.62)	(18.79)	(18.71)	(12.99)	(15.81)	(15.41)
Log-Likelihood	-7,989.60	-7,932.10	-7,933.90	-8026.20	-7,986.50	-7,991.70
Pseudo R^2	0.014	0.021	0.021	0.001	0.014	0.014
Observations	55,337	55,337	55,337	55,337	55,337	55,337

Table 4: **Logit regression of 8-K cyber incident report**

Logistic regression analyzing the term "cyber" occurrence in 8-K filings within 12 months after releasing a 10-K report and its associated cyber score. The dependent variable takes the value of 1 if the term "cyber" appears in the 8-Ks and 0 otherwise. The primary independent variable is the cyber score computed from the preceding 10-K report. The regression coefficient for the cyber score indicates its influence on the likelihood of "cyber" appearing in the subsequent 8-K filings.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**50.514*****	**53.229*****
	[46.66]	[65.46]
Firm Size (ln)	**0.008**	**0.039**
	[0.16]	[1.37]
Firm Age (ln)	**-0.346*****	**-0.492*****
	[-2.93]	[-8.80]
ROA	**0.027**	**0.014**
	[0.25]	[0.08]
Book to Market	**-0.028*****	**-0.138*****
	[-2.66]	[-5.04]
Market Beta	**-0.057***	**-0.115*****
	[-1.95]	[-2.83]
Intangibles/Assets	**-0.335***	**1.133*****
	[-1.75]	[5.51]
Debt/Assets	**-0.486****	**1.088*****
	[-2.13]	[2.59]
ROE	**-0.009**	**0.011**
	[-0.19]	[0.12]
Price/Earnings	**0.0003****	**0.00001**
	[2.10]	[0.04]
Profit Margin	**0.001**	**0.023*****
	[0.17]	[3.08]
Asset Turnover	**-0.056**	**-0.438*****
	[-0.66]	[-3.52]
Cash Ratio	**-0.0003**	**0.005**
	[-0.04]	[0.31]
Sales/Invested Capital	**0.013**	**0.134****
	[0.39]	[2.30]
Capital Ratio	**0.048**	**-2.200*****
	[0.25]	[-6.95]
R&D/Sales	**-0.005**	**-0.004**
	[-0.89]	[-0.42]
ROCE	**0.040**	**0.306***
	[0.43]	[1.92]
Readability	**0.090*****	**-0.164*****
	[2.81]	[-2.93]
Secret	**0.205***	**0.711*****
	[1.80]	[6.98]
Risk Length Table	**0.139*****	**0.254*****
	[4.67]	[7.36]
Volume per Cap.	**0.001**	**0.007*****
	[0.27]	[3.79]
Humans per Cap.	**-0.0004*****	**-0.000**
	[-11.02]	[-1.18]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.3193	0.2672

Table 5: **Determinants of firm-level overall cyber score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**43.921*****	**36.239*****
	[18.44]	[26.16]
Firm Size (ln)	**0.1046**	**0.2436*****
	[1.12]	[5.99]
Firm Age (ln)	**-0.6886*****	**-0.4258*****
	[-2.91]	[-5.15]
ROA	**-0.3729****	**-0.4243***
	[-2.07]	[-1.89]
Book to Market	**-0.0068**	**-0.0861****
	[-0.22]	[-2.07]
Market Beta	**-0.0759**	**0.0189**
	[-1.22]	[0.32]
Intangibles/Assets	**0.4**	**1.5571*****
	[1.02]	[5.33]
Debt/Assets	**0.1785**	**3.1032*****
	[0.37]	[5.77]
ROE	**-0.082**	**-0.0041**
	[-0.78]	[-0.03]
Price/Earnings	**-0.0002**	**-0.0003**
	[-0.89]	[-0.83]
Profit Margin	**-0.0059**	**0.0095**
	[-0.56]	[0.86]
Asset Turnover	**-0.0891**	**-0.1494**
	[-0.51]	[-0.8]
Cash Ratio	**0.0121**	**0.0217**
	[0.72]	[1.02]
Sales/Invested Capital	**-0.0733**	**-0.0916**
	[-1.14]	[-1.08]
Capital Ratio	**-0.12**	**-3.4501*****
	[-0.31]	[-8.31]
R&D/Sales	**-0.0225**	**-0.021**
	[-1.48]	[-1.48]
ROCE	**0.2894**	**0.4123***
	[1.47]	[1.74]
Readability	**0.1335**	**0.2959*****
	[1.1]	[2.85]
Secret	**0.4***	**0.5921*****
	[1.88]	[4.31]
Risk Length Table	**0.568*****	**0.7207*****
	[7.21]	[10.8]
Volume per Cap.	**-0.0044**	**-0.0029**
	[-0.9]	[-0.99]
Humans per Cap.	**-0.0014*****	**0.0005*****
	[-8.1]	[6.33]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.2221	0.2088

Table 6: **Determinants of firm-level cyber sentiment score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**48.709*****	**49.389*****
	[42.27]	[53.32]
Firm Size (ln)	**-0.014**	**0.0394**
	[-0.27]	[1.32]
Firm Age (ln)	**-0.5965*****	**-0.4996*****
	[-4.72]	[-8.04]
ROA	**0.0168**	**0.0293**
	[0.14]	[0.15]
Book to Market	**-0.0224****	**-0.1448*****
	[-1.97]	[-5.48]
Market Beta	**-0.0446**	**-0.1203*****
	[-1.45]	[-2.69]
Intangibles/Assets	**-0.3246**	**1.3698*****
	[-1.53]	[6.14]
Debt/Assets	**-0.626****	**0.8444***
	[-2.4]	[1.83]
ROE	**-0.0561**	**-0.0365**
	[-1.06]	[-0.39]
Price/Earnings	**0.0**	**-0.0003**
	[0.09]	[-0.96]
Profit Margin	**-0.0017**	**0.0259*****
	[-0.29]	[3.23]
Asset Turnover	**0.047**	**-0.5544*****
	[0.54]	[-4.03]
Cash Ratio	**0.0028**	**0.0129**
	[0.3]	[0.81]
Sales/Invested Capital	**-0.0259**	**0.1951*****
	[-0.7]	[3.06]
Capital Ratio	**0.1883**	**-2.2148*****
	[0.88]	[-6.28]
R&D/Sales	**-0.0063**	**0.0003**
	[-0.79]	[0.03]
ROCE	**0.097**	**0.3042***
	[0.93]	[1.73]
Readability	**0.0869****	**-0.1231****
	[2.47]	[-2.02]
Secret	**0.293****	**0.8431*****
	[2.47]	[7.77]
Risk Length Table	**0.118*****	**0.2876*****
	[3.6]	[6.68]
Volume per Cap.	**-0.0034**	**0.006*****
	[-1.05]	[2.66]
Humans per Cap.	**-0.0003*****	**-0.0001**
	[-9.34]	[-1.47]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.3224	0.2677

Table 7: **Determinants of firm-level command and data manipulation cyber score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**49.779*****	**50.925*****
	[41.1]	[57.92]
Firm Size (ln)	**-0.0834**	**-0.005**
	[-1.5]	[-0.16]
Firm Age (ln)	**-0.6247*****	**-0.5823*****
	[-4.97]	[-9.58]
ROA	**0.0446**	**0.0669**
	[0.4]	[0.34]
Book to Market	**-0.0085**	**-0.1247*****
	[-0.63]	[-3.88]
Market Beta	**-0.0529**	**-0.1494*****
	[-1.64]	[-3.4]
Intangibles/Assets	**-0.2478**	**1.0274*****
	[-1.2]	[4.62]
Debt/Assets	**-0.588****	**1.0181****
	[-2.33]	[2.29]
ROE	**0.008**	**0.0121**
	[0.15]	[0.13]
Price/Earnings	**0.0003***	**-0.0**
	[1.92]	[-0.04]
Profit Margin	**-0.0015**	**0.022*****
	[-0.26]	[2.97]
Asset Turnover	**-0.164***	**-0.5367*****
	[-1.83]	[-3.96]
Cash Ratio	**-0.0005**	**-0.0024**
	[-0.06]	[-0.15]
Sales/Invested Capital	**0.0404**	**0.1481****
	[1.17]	[2.42]
Capital Ratio	**0.2001**	**-2.1857*****
	[0.98]	[-6.42]
R&D/Sales	**-0.0093**	**-0.011**
	[-1.19]	[-1.18]
ROCE	**0.0148**	**0.2837**
	[0.15]	[1.61]
Readability	**0.1344*****	**-0.1081***
	[3.7]	[-1.78]
Secret	**0.2014***	**0.7763*****
	[1.65]	[7.15]
Risk Length Table	**0.144*****	**0.2837*****
	[4.1]	[7.02]
Volume per Cap.	**0.0006**	**0.008*****
	[0.12]	[3.59]
Humans per Cap.	**-0.0005*****	**-0.0**
	[-11.29]	[-0.64]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.3099	0.2564

Table 8: **Determinants of firm-level credential movement cyber score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**49.357*****	**52.308*****
	[47.79]	[65.49]
Firm Size (ln)	**-0.0327**	**0.0116**
	[-0.71]	[0.43]
Firm Age (ln)	**-0.1343**	**-0.4462*****
	[-1.2]	[-8.4]
ROA	**0.0298**	**0.1**
	[0.29]	[0.54]
Book to Market	**-0.0117**	**-0.1262*****
	[-0.96]	[-5.07]
Market Beta	**-0.0542***	**-0.127*****
	[-1.96]	[-3.29]
Intangibles/Assets	**-0.2192**	**0.9123*****
	[-1.23]	[4.78]
Debt/Assets	**-0.4056***	**1.0515*****
	[-1.85]	[2.6]
ROE	**0.0071**	**-0.0028**
	[0.16]	[-0.03]
Price/Earnings	**0.0001**	**-0.0001**
	[1.2]	[-0.29]
Profit Margin	**0.0004**	**0.0225*****
	[0.09]	[3.0]
Asset Turnover	**-0.0273**	**-0.4263*****
	[-0.36]	[-3.64]
Cash Ratio	**-0.0018**	**0.0015**
	[-0.2]	[0.11]
Sales/Invested Capital	**0.0107**	**0.1105****
	[0.34]	[2.03]
Capital Ratio	**-0.0628**	**-2.1721*****
	[-0.36]	[-7.0]
R&D/Sales	**-0.004**	**-0.0051**
	[-0.65]	[-0.54]
ROCE	**-0.0611**	**0.1708**
	[-0.72]	[1.12]
Readability	**0.1256*****	**-0.1214****
	[3.34]	[-2.15]
Secret	**0.059**	**0.6653*****
	[0.55]	[6.89]
Risk Length Table	**0.0929*****	**0.1794*****
	[3.29]	[5.55]
Volume per Cap.	**-0.0007**	**0.0056****
	[-0.21]	[2.29]
Humans per Cap.	**-0.0002*****	**-0.0**
	[-7.49]	[-0.82]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.2193	0.1566

Table 9: **Determinants of firm-level persistence and evasion cyber score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

Dependent variable: Firm-level indicator of cyber score		
	Model 1	Model 2
Constant	**48.897*****	**49.466*****
	[42.14]	[56.36]
Firm Size (ln)	**0.0069**	**0.0873*****
	[0.13]	[2.99]
Firm Age (ln)	**-0.7352*****	**-0.523*****
	[-5.92]	[-9.11]
ROA	**-0.0807**	**-0.1396**
	[-0.76]	[-0.78]
Book to Market	**-0.0347*****	**-0.1499*****
	[-2.75]	[-5.58]
Market Beta	**-0.0577***	**-0.0962****
	[-1.92]	[-2.34]
Intangibles/Assets	**-0.1576**	**1.5511*****
	[-0.73]	[7.31]
Debt/Assets	**-0.4222***	**1.6297*****
	[-1.67]	[3.81]
ROE	**0.0376**	**0.0192**
	[0.7]	[0.21]
Price/Earnings	**0.0002**	**0.0**
	[1.6]	[0.05]
Profit Margin	**-0.0009**	**0.0191*****
	[-0.21]	[2.63]
Asset Turnover	**-0.033**	**-0.4303*****
	[-0.36]	[-3.38]
Cash Ratio	**0.0055**	**0.015**
	[0.56]	[0.96]
Sales/Invested Capital	**0.0072**	**0.146****
	[0.21]	[2.46]
Capital Ratio	**0.0664**	**-2.5668*****
	[0.33]	[-7.98]
R&D/Sales	**-0.0094**	**-0.0015**
	[-1.59]	[-0.15]
ROCE	**0.0523**	**0.3945****
	[0.52]	[2.55]
Readability	**0.0651**	**-0.1943*****
	[1.64]	[-3.14]
Secret	**0.2292***	**0.6711*****
	[1.95]	[6.58]
Risk Length Table	**0.1945*****	**0.3505*****
	[5.87]	[9.16]
Volume per Cap.	**-0.0021**	**0.0036**
	[-0.53]	[1.45]
Humans per Cap.	**-0.0005*****	**0.0**
	[-10.84]	[0.44]
Year fixed effect	Yes	Yes
Industry fixed effect	No	Yes
Firm fixed effect	Yes	No
Observations	25531	25531
R^2 within	0.4331	0.3920

Table 10: **Determinants of firm-level preparation and reconnaissance cyber score**

This table reports the results of cyber score regressions on firm characteristics. Year-, industry-, and firm-fixed effects are controlled. T-statistics are reported in brackets. The variables are standardized, and the standard errors are clustered at the firm level. $*$, $**$, and $***$ indicate significance at the 10%, 5% and 1% levels, respectively. All characteristics are defined in Table A1.

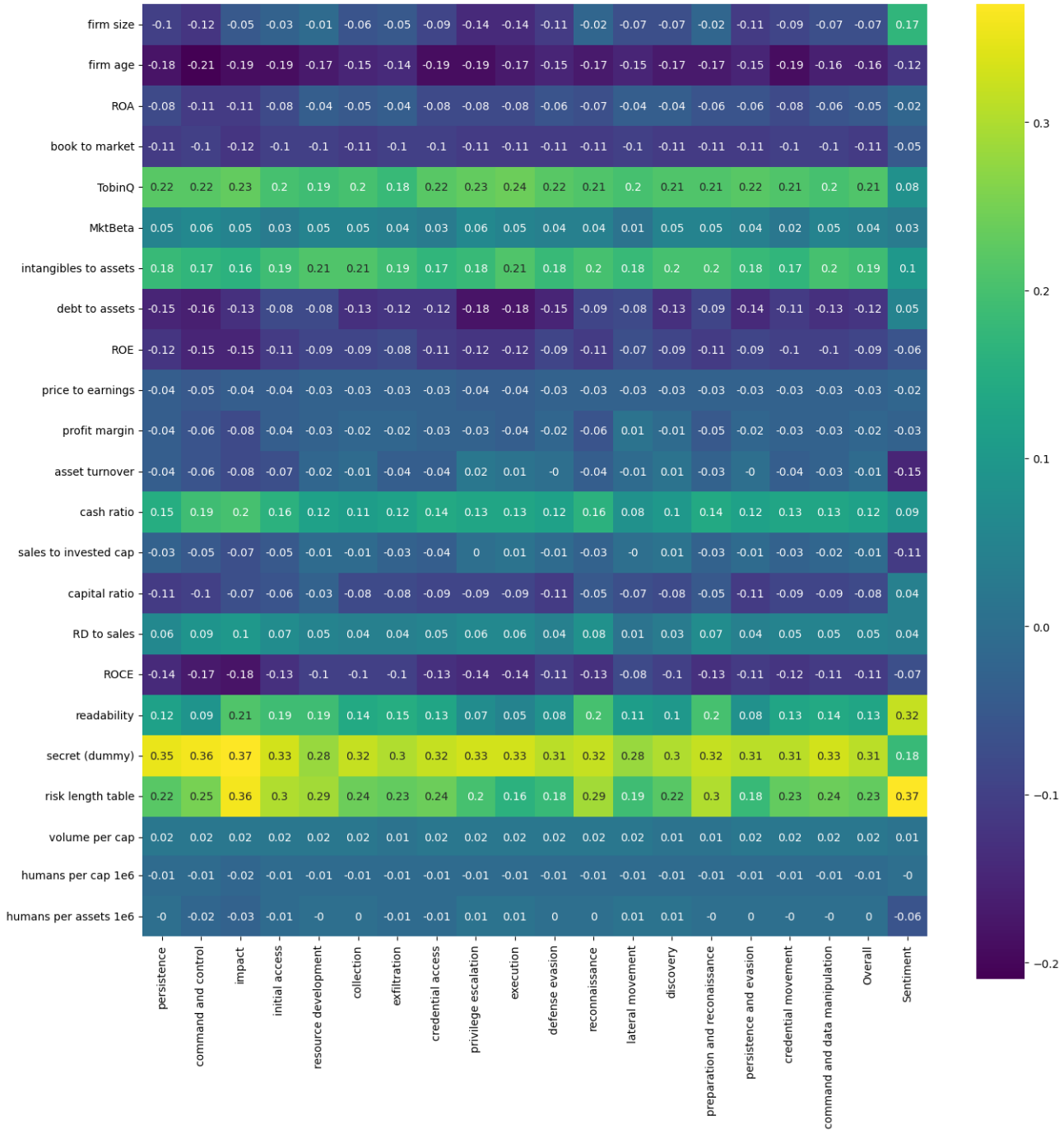

Fig. 14: **Correlations of all cyber scores with financial characteristics**

Firm-wise correlations of the sub-cyber scores of the 14 MITRE ATT&CK tactics, the four aggregated sub-cyber scores of the super-tactics, as well as the overall cyber score and the cyber sentiment score with the financial characteristics of the firms.

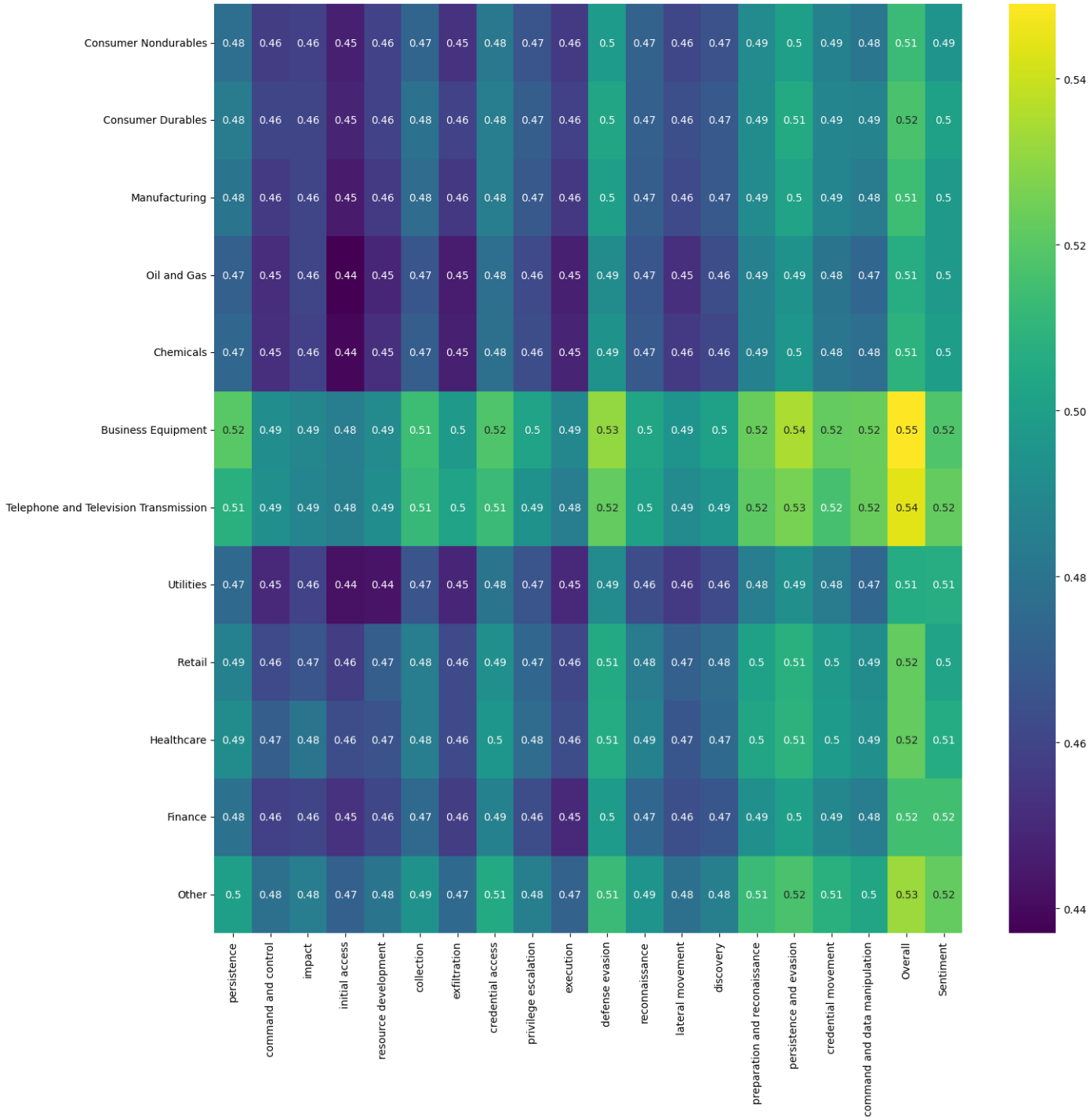

Fig. 15: **Average cyber score across industries**

The respective average cyber scores of each firm (from 2009-Q1 to 2023-Q4) are computed and averaged across the industry the firms belong to, thus obtaining the different averaged cyber scores aggregate for each industry. Firms are classified into industries using the Fama-French 12 industry classification.

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.82*** **	**0.93*** **	**1.04*** **	**1.22*** **	**1.44*** **	**0.62** **
	[3.27]	[3.46]	[3.65]	[4.65]	[4.54]	[2.05]
CAPM alpha	**-0.18**	**-0.12**	**-0.08**	**0.14**	**0.36** **	**0.54**
	[-0.85]	[-0.93]	[-0.84]	[1.47]	[2.14]	[1.49]
FFC alpha	**-0.09**	**-0.05**	**0.0**	**0.15***	**0.27*** **	**0.36** **
	[-0.88]	[-0.57]	[0.04]	[1.71]	[3.04]	[2.2]
FF5 alpha	**-0.14**	**-0.1**	**0.0**	**0.13**	**0.29*** **	**0.44*** **
	[-1.57]	[-1.18]	[0.01]	[1.47]	[3.16]	[2.88]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.49	0.51	0.52	0.53	0.57	-
Sharp Ratio	0.61	0.69	0.72	0.88	1.02	0.68

Table 11: **Average monthly excess returns and alphas (in percent) using the overall cyber score**

FFC refers to the four-factor model of Carhart (1997), and FF5 refers to the five-factor model of Fama and French (2015). Panel B shows the average number of firms in each portfolio and the average cyber risk of the portfolios. T-statistics are reported in brackets. $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% levels, respectively. The time ranges from January 2009 to December 2023.

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.99*** **	**1.08*** **	**1.24*** **	**1.15*** **	**1.14*** **	**0.14**
	[3.92]	[4.61]	[4.78]	[3.94]	[3.88]	[1.21]
CAPM alpha	**-0.03**	**0.04**	**0.19***	**0.02**	**0.05**	**0.08**
	[-0.27]	[0.31]	[1.91]	[0.31]	[0.59]	[0.58]
FFC alpha	**0.0**	**0.05**	**0.17***	**0.04**	**0.05**	**0.05**
	[0.02]	[0.48]	[1.78]	[0.59]	[0.66]	[0.45]
FF5 alpha	**-0.03**	**-0.02**	**0.12**	**0.07**	**0.1**	**0.12**
	[-0.41]	[-0.18]	[1.31]	[1.15]	[1.17]	[1.15]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.46	0.49	0.51	0.53	0.57	-
Sharp Ratio	0.75	0.8	0.92	0.81	0.82	0.3

Table 12: **Average monthly excess returns and alphas (in percent) using the cyber sentiment score**

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.81*** **	**0.91*** **	**1.12*** **	**1.17*** **	**1.46*** **	**0.65***
	[3.18]	[3.35]	[4.31]	[4.55]	[4.33]	[1.94]
CAPM alpha	**-0.22**	**-0.13**	**0.06**	**0.05**	**0.39** **	**0.6**
	[-0.94]	[-1.07]	[0.75]	[0.61]	[2.01]	[1.49]
FFC alpha	**-0.11**	**-0.07**	**0.1** **	**0.04**	**0.3*** **	**0.42** **
	[-1.0]	[-0.77]	[2.18]	[0.5]	[2.69]	[2.12]
FF5 alpha	**-0.16**	**-0.12**	**0.11** **	**0.03**	**0.33*** **	**0.49*** **
	[-1.66]	[-1.35]	[2.01]	[0.37]	[2.76]	[2.61]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.46	0.48	0.49	0.5	0.54	-
Sharp Ratio	0.59	0.68	0.82	0.82	1.04	0.71

Table 13: **Average monthly excess returns and alphas (in percent) using the command and data manipulation cyber score**

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.88*** **	**0.9*** **	**1.04*** **	**1.13*** **	**1.49*** **	**0.61** **
	[3.63]	[3.51]	[3.5]	[4.34]	[4.66]	[2.06]
CAPM alpha	**-0.12**	**-0.14**	**-0.1**	**0.07**	**0.4** **	**0.52**
	[-0.6]	[-1.01]	[-1.16]	[1.11]	[2.31]	[1.48]
FFC alpha	**-0.04**	**-0.07**	**-0.02**	**0.09**	**0.3*** **	**0.34** **
	[-0.41]	[-0.76]	[-0.29]	[1.4]	[3.09]	[2.07]
FF5 alpha	**-0.1**	**-0.11**	**-0.02**	**0.08**	**0.33*** **	**0.43*** **
	[-1.15]	[-1.25]	[-0.32]	[1.23]	[3.19]	[2.74]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.46	0.48	0.49	0.51	0.54	-
Sharp Ratio	0.66	0.67	0.72	0.83	1.04	0.69

Table 14: **Average monthly excess returns and alphas (in percent) using the credential movement cyber score**

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.84*** **	**0.95*** **	**1.04*** **	**1.11*** **	**1.49*** **	**0.64** **
	[3.35]	[3.61]	[3.75]	[4.44]	[4.54]	[2.01]
CAPM alpha	**-0.18**	**-0.08**	**-0.03**	**0.05**	**0.39** **	**0.58**
	[-0.86]	[-0.54]	[-0.39]	[0.66]	[2.16]	[1.53]
FFC alpha	**-0.1**	**0.01**	**0.03**	**0.08**	**0.29*** **	**0.39** **
	[-0.89]	[0.13]	[0.38]	[1.47]	[2.9]	[2.23]
FF5 alpha	**-0.15**	**-0.05**	**0.03**	**0.05**	**0.33*** **	**0.48*** **
	[-1.66]	[-0.78]	[0.5]	[0.78]	[3.09]	[2.97]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.48	0.49	0.5	0.52	0.55	-
Sharp Ratio	0.61	0.71	0.76	0.82	1.03	0.67

Table 15: **Average monthly excess returns and alphas (in percent) using the persistence and evasion cyber score**

	P1	P2	P3	P4	P5	P5-P1
A. Portfolios sorted by cyber score						
avg. excess ret.	**0.86*** **	**0.85*** **	**1.15*** **	**1.11*** **	**1.43*** **	**0.57** **
	[3.54]	[3.13]	[4.22]	[3.94]	[4.69]	[1.97]
CAPM alpha	**-0.09**	**-0.23**	**0.02**	**0.02**	**0.37** **	**0.46**
	[-0.44]	[-1.97]	[0.2]	[0.29]	[2.39]	[1.34]
FFC alpha	**-0.02**	**-0.16**	**0.07**	**0.04**	**0.28*** **	**0.3***
	[-0.14]	[-2.45]	[1.23]	[0.56]	[3.07]	[1.75]
FF5 alpha	**-0.09**	**-0.19**	**0.11***	**0.01**	**0.3*** **	**0.38** **
	[-0.95]	[-3.08]	[1.76]	[0.19]	[3.17]	[2.42]
B. Characteristics						
Nb. firms	628.48	629.1	629.01	629.1	629.67	-
Avg. cyber score	0.47	0.48	0.5	0.51	0.54	-
Sharp Ratio	0.67	0.61	0.8	0.8	1.03	0.69

Table 16: **Average monthly excess returns and alphas (in percent) using the preparation and reconnaissance cyber score**

	Q1	Q2	Q3	Q4	Q5		Q1	Q2	Q3	Q4	Q5		Q1	Q2	Q3	Q4	Q5
Double sorted portfolios with overall cyber score																	
Beta Q1	1.01	1.01	1.13	1.37	1.42	BM Q1*	1.09	0.98	1.14	1.18	1.24	Size Q1	0.86	0.96	1.06	1.22	1.32
Beta Q2	0.89	1.00	1.07	1.25	1.33	BM Q2	0.86	0.92	1.01	1.25	1.37	Size Q2	0.87	0.96	1.05	1.25	1.32
Beta Q3	0.89	1.02	1.01	1.22	1.28	BM Q3	0.85	1.01	1.04	1.23	1.33	Size Q3	0.87	0.95	1.05	1.25	1.32
Beta Q4	0.84	0.90	1.04	1.23	1.25	BM Q4	0.83	0.99	1.03	1.22	1.35	Size Q4	0.86	0.95	1.08	1.26	1.33
Beta Q5	0.83	0.92	1.02	1.21	1.30	BM Q5	0.87	0.95	1.07	1.22	1.33	Size Q5	1.09	1.15	1.20	1.19	1.39
Double sorted portfolios with cyber sentiment score																	
Beta Q1	1.16	1.23	1.27	1.25	1.09	BM Q1*	1.13	1.15	1.04	1.12	1.12	Size Q1*	0.97	1.20	1.10	1.15	1.07
Beta Q2*	0.95	1.00	1.36	1.18	1.07	BM Q2*	0.93	1.01	1.30	1.13	1.09	Size Q2*	0.96	1.12	1.21	1.13	1.07
Beta Q3*	1.06	1.00	1.07	1.11	1.10	BM Q3*	0.96	1.13	1.22	1.15	1.08	Size Q3*	0.96	1.11	1.19	1.13	1.07
Beta Q4*	0.95	1.08	1.10	1.09	1.03	BM Q4*	0.95	1.09	1.21	1.14	1.08	Size Q4*	0.95	1.03	1.24	1.11	1.08
Beta Q5*	0.96	1.00	1.20	1.13	1.04	BM Q5*	0.95	1.09	1.20	1.15	1.05	Size Q5*	1.23	1.16	1.20	1.17	1.29
Double sorted portfolios with command and data manipulation cyber score																	
Beta Q1	1.04	1.06	1.18	1.27	1.41	BM Q1*	1.07	0.94	1.22	1.15	1.24	Size Q1	0.90	0.89	1.16	1.13	1.33
Beta Q2	0.89	0.93	1.14	1.19	1.36	BM Q2	0.83	0.89	1.11	1.16	1.39	Size Q2*	0.89	0.91	1.16	1.12	1.33
Beta Q3*	0.91	0.90	1.16	1.09	1.34	BM Q3	0.89	0.90	1.17	1.16	1.34	Size Q3	0.87	0.91	1.16	1.14	1.34
Beta Q4	0.85	0.85	1.17	1.14	1.25	BM Q4	0.85	0.93	1.15	1.20	1.32	Size Q4	0.85	0.96	1.11	1.21	1.34
Beta Q5	0.87	0.85	1.11	1.12	1.30	BM Q5	0.86	0.96	1.13	1.20	1.31	Size Q5	1.11	1.19	1.16	1.16	1.41
Double sorted portfolios with credential movement cyber score																	
Beta Q1*	1.05	0.96	1.16	1.22	1.49	BM Q1*	1.03	1.03	1.18	1.09	1.25	Size Q1	0.90	0.93	1.06	1.10	1.39
Beta Q2*	0.93	0.94	1.08	1.03	1.42	BM Q2	0.86	0.90	1.07	1.16	1.41	Size Q2	0.90	0.92	1.09	1.09	1.38
Beta Q3	0.90	0.99	1.03	1.08	1.37	BM Q3	0.93	0.92	1.05	1.09	1.41	Size Q3	0.90	0.92	1.08	1.08	1.40
Beta Q4	0.85	0.88	1.09	1.09	1.31	BM Q4	0.90	0.90	1.07	1.09	1.41	Size Q4	0.89	0.93	1.03	1.10	1.42
Beta Q5	0.90	0.87	1.04	1.11	1.35	BM Q5	0.88	0.95	1.06	1.06	1.41	Size Q5	1.13	1.13	1.17	1.22	1.39
Double sorted portfolios with persistence and evasion cyber score																	
Beta Q1*	1.02	1.07	1.12	1.25	1.46	BM Q1*	1.06	1.01	1.18	1.19	1.19	Size Q1	0.88	0.99	1.08	1.12	1.36
Beta Q2*	0.88	1.06	1.05	1.13	1.40	BM Q2*	0.85	1.03	0.98	1.19	1.41	Size Q2	0.89	1.00	1.08	1.15	1.34
Beta Q3	0.90	1.05	1.06	1.09	1.33	BM Q3	0.90	1.00	1.03	1.15	1.38	Size Q3	0.88	0.99	1.05	1.14	1.38
Beta Q4	0.86	0.97	1.05	1.11	1.29	BM Q4	0.87	1.00	1.03	1.17	1.38	Size Q4	0.86	1.02	1.05	1.14	1.39
Beta Q5	0.87	0.91	0.99	1.18	1.32	BM Q5	0.86	1.00	1.05	1.16	1.38	Size Q5	1.17	1.10	1.20	1.16	1.39
Double sorted portfolios with preparation and reconnaissance cyber score																	
Beta Q1*	1.10	0.91	1.28	1.21	1.41	BM Q1*	1.08	0.99	1.12	1.13	1.24	Size Q1*	0.91	0.84	1.13	1.15	1.34
Beta Q2*	0.93	0.87	1.20	1.13	1.35	BM Q2*	0.85	0.78	1.18	1.19	1.37	Size Q2*	0.92	0.83	1.14	1.20	1.31
Beta Q3	0.92	0.90	1.13	1.11	1.29	BM Q3*	0.90	0.93	1.15	1.09	1.34	Size Q3*	0.90	0.87	1.17	1.13	1.32
Beta Q4*	0.90	0.77	1.20	1.07	1.27	BM Q4	0.86	0.89	1.15	1.12	1.34	Size Q4*	0.86	0.86	1.20	1.12	1.34
Beta Q5*	0.89	0.78	1.13	1.12	1.29	BM Q5	0.85	0.90	1.13	1.14	1.34	Size Q5	1.13	1.15	1.19	1.19	1.38

Table 17: **Average returns of the double sorted portfolios**

Q1 to **Q5** represent quintiles. The sorting of firms is done according to market beta (Beta), book-to-market ratios (BM), or firm size (Size) and then on the relevant cyber score. The average returns are given in percent. $*$ indicates that the returns are not increasing monotonically with the quintile of the cyber score (with an incertitude of -0.03%).

	M.1	M.2	M.3	M.4	M.5
Market	**0.011*****		**0.009****	**0.013*****	**0.009****
	[2.886]		[2.526]	[3.507]	[2.429]
Cyber		**0.054***	**0.051***	**0.051****	**0.04**
		[1.925]	[1.807]	[2.097]	[1.547]
HML				**0.003**	**0.003**
				[1.176]	[0.964]
SMB				**-0.001**	**0.001**
				[-0.223]	[0.636]
UMD				**0.002**	
				[0.766]	
CMA					**-0.001**
					[-0.627]
RMW					**0.002**
					[0.776]
Constant	**0.001**	**-0.017**	**-0.024**	**-0.029****	**-0.019**
	[0.148]	[-1.083]	[-1.586]	[-2.223]	[-1.357]
$\overline{R^2_{\text{adj}}}$	0.067	0.158	0.22	0.296	0.309
MAPE	0.013	0.012	0.012	0.01	0.009

Table 18: **Fama-MacBeth for overall cyber score**

This table reports the results of Fama-MacBeth regressions of 20 value-weighted portfolios sorted on their cyber score. These portfolios are regressed each month on portfolio value-weighted betas with the market, HML, SMB, MOM, RMW, and CMA. "Cyber" is the value-weighted cyber score of each portfolio. HML and SMB refer to the book-to-market and size factors from Fama and French (1992). UMD refers to the momentum factor from Carhart (1997). CMA and RMW refer to the investment and operating profitability factors from Fama and French (2015). $\overline{R^2_{\text{adj}}}$ is the average adjusted R-squared, and MAPE is the mean average pricing error (mean average of the absolute value of the residuals). 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 2023.

	M.1	M.2	M.3	M.4	M.5
Market	**0.005**		**0.004**	**0.005**	**0.005**
	[1.335]		[0.951]	[1.308]	[1.27]
Cyber		**0.003**	**0.0**	**0.003**	**-0.008**
		[0.277]	[0.003]	[0.228]	[-0.501]
HML				**0.001**	**-0.001**
				[0.381]	[-0.235]
SMB				**0.001**	**0.001**
				[0.304]	[0.383]
UMD				**0.005**	
				[1.508]	
CMA					**-0.001**
					[-0.488]
RMW					**0.001**
					[0.486]
Constant	**0.006**	**0.01**	**0.008**	**0.005**	**0.011**
	[1.643]	[1.607]	[1.209]	[0.754]	[1.439]
$\overline{R^2_{\text{adj}}}$	0.077	0.042	0.108	0.202	0.237
MAPE	0.013	0.013	0.012	0.011	0.01

Table 19: **Fama-McBeth for cyber sentiment score**

	M.1	M.2	M.3	M.4	M.5
Market	**0.009****		**0.007***	**0.009****	**0.01****
	[2.328]		[1.716]	[2.493]	[2.574]
Cyber		**0.044***	**0.051***	**0.056****	**0.038**
		[1.673]	[1.914]	[2.335]	[1.401]
HML				**0.003**	**0.003**
				[1.122]	[0.942]
SMB				**0.002**	**0.002**
				[0.884]	[0.692]
UMD				**0.0**	
				[0.005]	
CMA					**-0.001**
					[-0.35]
RMW					**0.001**
					[0.561]
Constant	**0.003**	**-0.01**	**-0.02**	**-0.026****	**-0.016**
	[0.661]	[-0.733]	[-1.439]	[-2.032]	[-1.141]
$\overline{R^2_{\text{adj}}}$	0.053	0.154	0.206	0.315	0.322
MAPE	0.014	0.013	0.012	0.01	0.01

Table 20: **Fama-McBeth for command and data manipulation cyber score**

	M.1	M.2	M.3	M.4	M.5
Market	**0.007****		**0.004**	**0.009****	**0.007***
	[2.106]		[1.036]	[2.454]	[1.894]
Cyber		**0.051***	**0.056****	**0.065****	**0.056****
		[1.906]	[2.101]	[2.419]	[2.062]
HML				**0.005***	**0.003**
				[1.67]	[1.162]
SMB				**-0.001**	**0.001**
				[-0.379]	[0.46]
UMD				**-0.001**	
				[-0.243]	
CMA					**-0.002**
					[-1.145]
RMW					**0.001**
					[0.458]
Constant	**0.004**	**-0.014**	**-0.02**	**-0.03****	**-0.023**
	[1.04]	[-0.982]	[-1.487]	[-2.228]	[-1.649]
$\overline{R^2_{\text{adj}}}$	0.057	0.157	0.219	0.308	0.313
MAPE	0.013	0.013	0.012	0.01	0.009

Table 21: **Fama-McBeth for credential movement cyber score**

	M.1	M.2	M.3	M.4	M.5
Market	**0.004**		**0.002**	**0.004**	**0.003**
	[1.201]		[0.472]	[1.222]	[0.801]
Cyber		**0.056***	**0.061****	**0.073****	**0.071****
		[1.835]	[2.062]	[2.451]	[2.193]
HML				**0.003**	**0.004**
				[1.255]	[1.357]
SMB				**0.001**	**0.001**
				[0.442]	[0.474]
UMD				**-0.002**	
				[-0.477]	
CMA					**-0.0**
					[-0.192]
RMW					**0.002**
					[1.370]
Constant	**0.007**	**-0.017**	**-0.021**	**-0.03***	**-0.028**
	[1.644]	[-1.04]	[-1.349]	[-1.944]	[-1.641]
$\overline{R^2_{\text{adj}}}$	0.058	0.167	0.215	0.3	0.304
MAPE	0.013	0.012	0.012	0.01	0.009

Table 22: **Fama-McBeth for persistence and evasion cyber score**

	M.1	M.2	M.3	M.4	M.5
Market	**0.009***		**0.006**	**0.011****	**0.01****
	[1.951]		[1.531]	[2.541]	[2.055]
Cyber		**0.047***	**0.046***	**0.049****	**0.052****
		[1.848]	[1.888]	[2.262]	[2.299]
HML				**0.003**	**0.004**
				[1.053]	[1.52]
SMB				**0.001**	**0.001**
				[0.233]	[0.418]
UMD				**0.001**	
				[0.168]	
CMA					**-0.001**
					[-0.466]
RMW					**0.002**
					[0.862]
Constant	**0.003**	**-0.012**	**-0.018**	**-0.024****	**-0.025****
	[0.674]	[-0.908]	[-1.345]	[-2.047]	[-2.008]
$\overline{R^2_{\text{adj}}}$	0.072	0.137	0.201	0.274	0.288
MAPE	0.014	0.013	0.012	0.01	0.01

Table 23: **Fama-McBeth for preparation and reconaissance cyber score**

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.451	0.107	0.868	0.737	0.783	0.876
FF5 + CyberFactor	1.088	0.367	0.888	0.833	0.671	0.877
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.495	0.090	0.806	1.021	0.441	0.895

Table 24: **GRS test for overall cyber score**

This table reports the results of time series regressions of 20 value-weighted portfolios (sorted on the cyber score, the size of firms, the market beta or the book-to-market ratio) on the five-factor model ofFama and French (2015) (FF5) and the "CyberFactor", *i.e.* the factor built as the long-short of extreme quintile portfolios sorted on the relevant cyber score (P5-P1). The p-value is the probability that the alphas of the 20 regressions are jointly zero. A probability lower than 10% means that the hypothesis that alphas are jointly zero can be rejected at the 10% level. The study period is from January 2009 to December 2023.

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.254	0.218	0.854	0.737	0.783	0.876
FF5 + CyberFactor	1.198	0.263	0.864	0.748	0.771	0.877
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.488	0.093	0.796	1.188	0.271	0.892

Table 25: **GRS test for cyber sentiment score**

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.558	0.070	0.855	0.737	0.783	0.876
FF5 + CyberFactor	1.119	0.335	0.877	0.844	0.657	0.877
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.472	0.099	0.807	1.026	0.436	0.895

Table 26: **GRS test for command and data manipulation cyber score**

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.539	0.076	0.864	0.737	0.783	0.876
FF5 + CyberFactor	1.192	0.268	0.884	0.770	0.746	0.877
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.441	0.111	0.804	0.997	0.469	0.895

Table 27: **GRS test for credential movement**

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.465	0.101	0.868	0.737	0.783	0.876
FF5 + CyberFactor	1.128	0.326	0.890	0.884	0.608	0.878
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.500	0.088	0.806	1.021	0.441	0.896

Table 28: **GRS test for persistence and evasion cyber score**

	GRS	p-value	$\overline{R2}$	GRS	p-value	$\overline{R2}$
	Sorted on cyber score			Sorted on size		
FF5	1.516	0.083	0.861	0.737	0.783	0.876
FF5 + CyberFactor	1.216	0.248	0.879	0.807	0.703	0.877
	Sorted on market beta			Sorted on book-to-market		
FF5	1.541	0.075	0.793	1.240	0.229	0.891
FF5 + CyberFactor	1.546	0.076	0.807	0.990	0.477	0.896

Table 29: **GRS test for preparation and reconaissance cyber score**

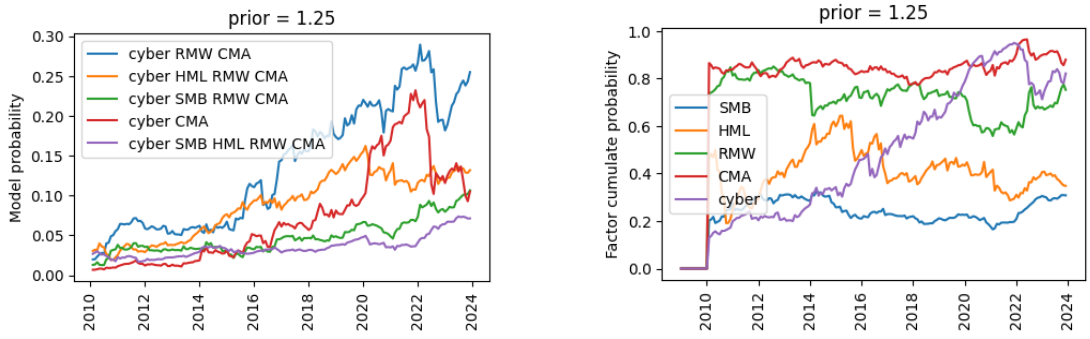

Fig. 16: **Factor model posterior probabilities using overall cyber score**

The first figure depicts the probabilities of being a better set of pricing factors for the shown subset compared to all possible subsets of factors. I present only the top five models, ranked by the probability at the end of the sample, meaning that all other subsets have lower pricing abilities than the ones presented here. 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). "cyber" refers to the long-short portfolio built on the cyber score of interest (P5-P1). The prior multiple is 1.25, and the study period is from January 2010 to December 2022. The second figure shows the cumulative probabilities, *i.e.* the sum of probabilities of all the pricing subsets containing the factor on a similar time range.

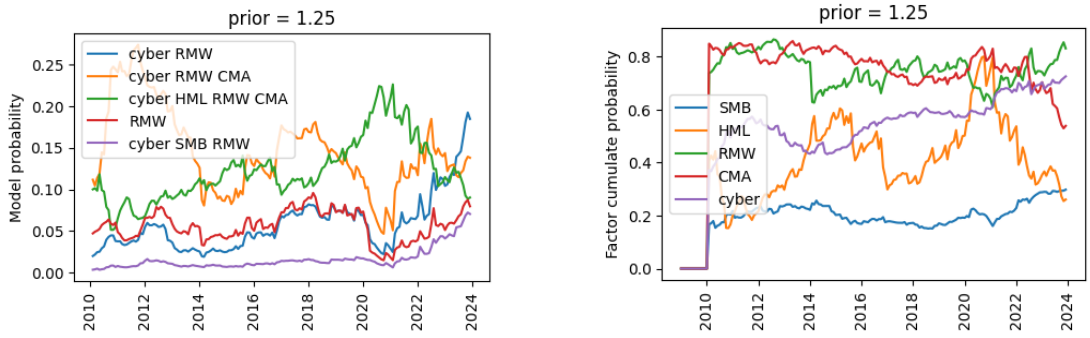

Fig. 17: **Factor model posterior probabilities using cyber sentiment score**

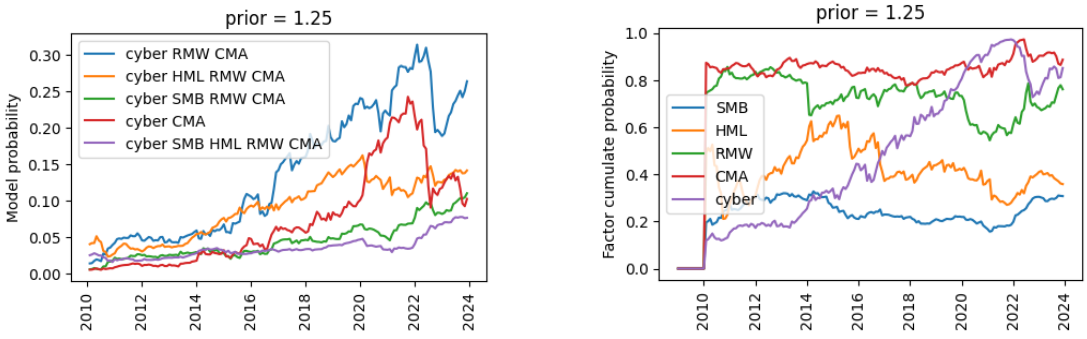

Fig. 18: **Factor model posterior probabilities using command and data manipulation cyber score**

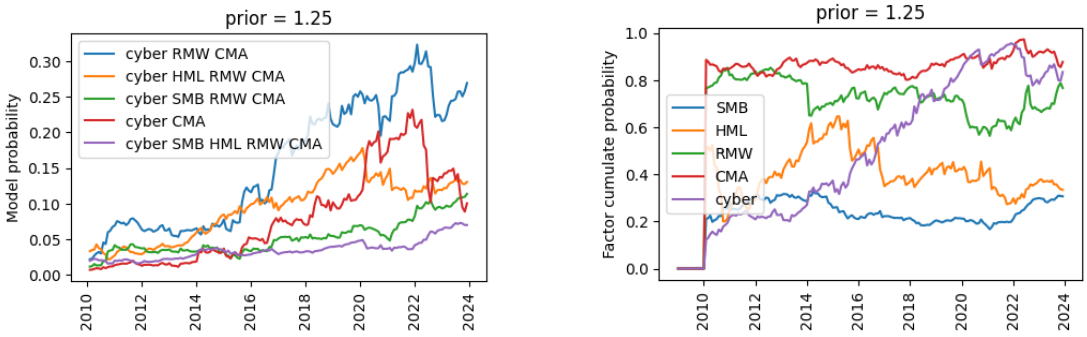

Fig. 19: **Factor model posterior probabilities using credential movement cyber score**

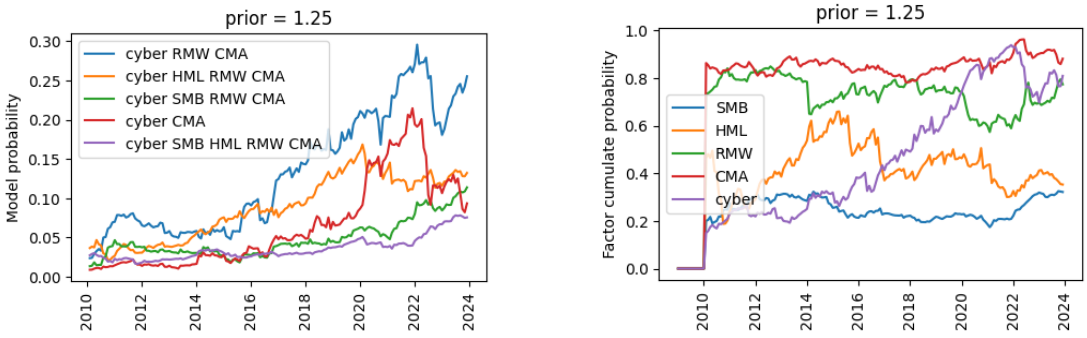

Fig. 20: **Factor model posterior probabilities using persistence and evasion cyber score**

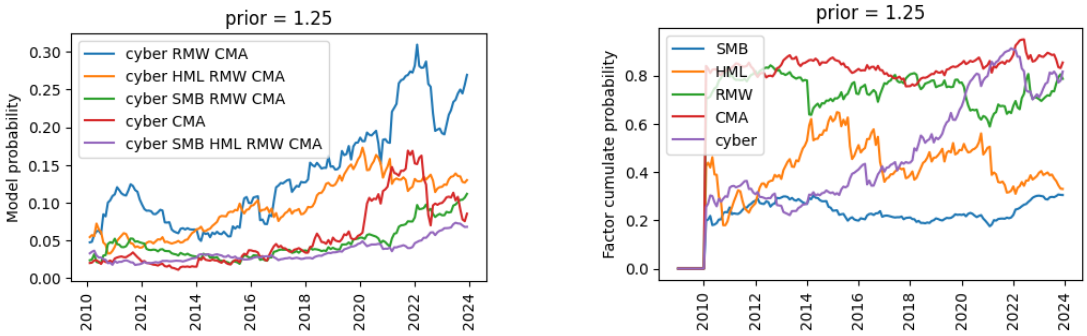

Fig. 21: **Factor model posterior probabilities using preparation and reconnaissance cyber score**

Cyber score	P5 (top 20%)		P20 (top 5%)	
	t-stat.	p-value	t-stat.	p-value
persistence	-0.0372	0.9703	-0.0768	0.9388
command and control	0.0605	0.9518	0.7315	0.4649
impact	0.0692	0.9449	0.0011	0.9992
initial access	-0.0751	0.9402	0.1292	0.8972
resource development	0.3001	0.7643	0.2834	0.7770
collection	-0.0099	0.9921	0.0890	0.9291
exfiltration	0.0041	0.9967	0.4411	0.6593
credential access	-0.0167	0.9867	0.3022	0.7626
privilege escalation	-0.0461	0.9632	0.1641	0.8697
execution	0.2384	0.8117	0.1040	0.9172
defense evasion	0.0474	0.9622	-0.1037	0.9174
reconnaissance	0.1565	0.8757	0.2418	0.8091
lateral movement	-0.0665	0.9470	-0.2061	0.8368
discovery	0.0419	0.9666	0.1053	0.9162
preparation and reconnaissance	0.1050	0.9165	0.2395	0.8108
persistence and evasion	-0.1185	0.9058	-0.0888	0.9293
credential movement	0.0242	0.9807	-0.0176	0.9860
command and data manipulation	0.0894	0.9288	0.0601	0.9521
sentiment	0.6125	0.5405	1.1653	0.2446

Table 30: **Cyber based portfolio returns differences**

The table displays the outcomes of Welch's t-test, the statistical method used to evaluate the significance of mean differences with the possibility of different variances, applied to each cyber score time series in comparison to the overall cyber score time series. These time series are the monthly returns of cyber-based portfolios: P5 (constructed with 5 quantiles, taking the top 20%) and P20 (constructed with 20 quantiles, taking the top 5%).

	P1	P2	P3	P4	P5	P5-P1
CAR[-1,1]	-0.146	0.001	-0.021	-0.103	0.206	0.352
t-statistic	-0.311	0.002	-0.058	-0.342	0.616	0.450
CAR[-1,3]	-0.197	-0.040	-0.178	-0.051	0.194	0.390
t-statistic	-0.540	-0.115	-0.626	-0.220	0.748	0.644

Table 31: **Cumulative abnormal returns of cyber-based portfolios**

To estimate the cumulative abnormal returns (CAR), I use the market model around December 14, 2020, as t=0. Note that it was a Monday. Therefore, $t = -1$ corresponds to Friday, December 11. The beta of the market model is set up thanks to the returns of the prior year. The abnormal returns are given in percent. The portfolios are based on the overall cyber score.

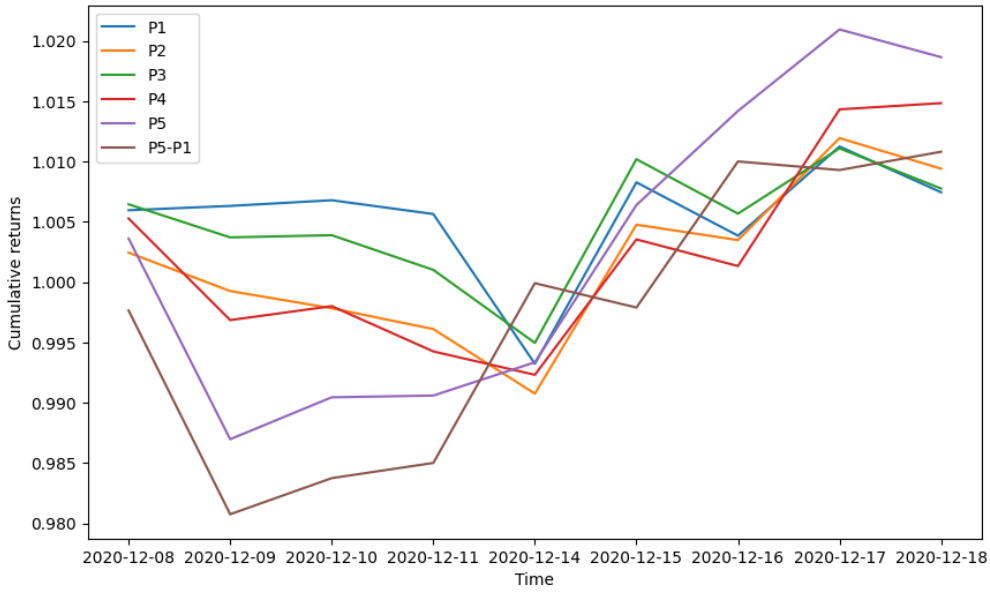

Fig. 22: **Cumulative returns of cyber-based portfolio around SolarWinds breach**

Evolution of the portfolio based on the overall cyber score if 1 dollar was invested the December 7, 2020. Note that the closed trading days do not appear.

	CAR[-1,1]	CAR[-1,3]
overall	0.078	0.055
	[0.151]	[0.138]
preparation and reconnaissance	0.036	0.146
	[0.082]	[0.429]
persistence and evasion	0.163	0.228
	[0.301]	[0.543]
credential movement	0.141	0.312
	[0.255]	[0.724]
command and data manipulation	0.207	0.227
	[0.427]	[0.603]

Table 32: **Cumulative abnormal returns of cyber-based P20**

To estimate the cumulative abnormal returns (CAR), I use the market model around December 14, 2020, as $t = 0$. The beta of the market model is set up thanks to the returns of the prior year. The abnormal returns are given in percent. The t-statistics associated with the abnormal returns are given in the parenthesis.

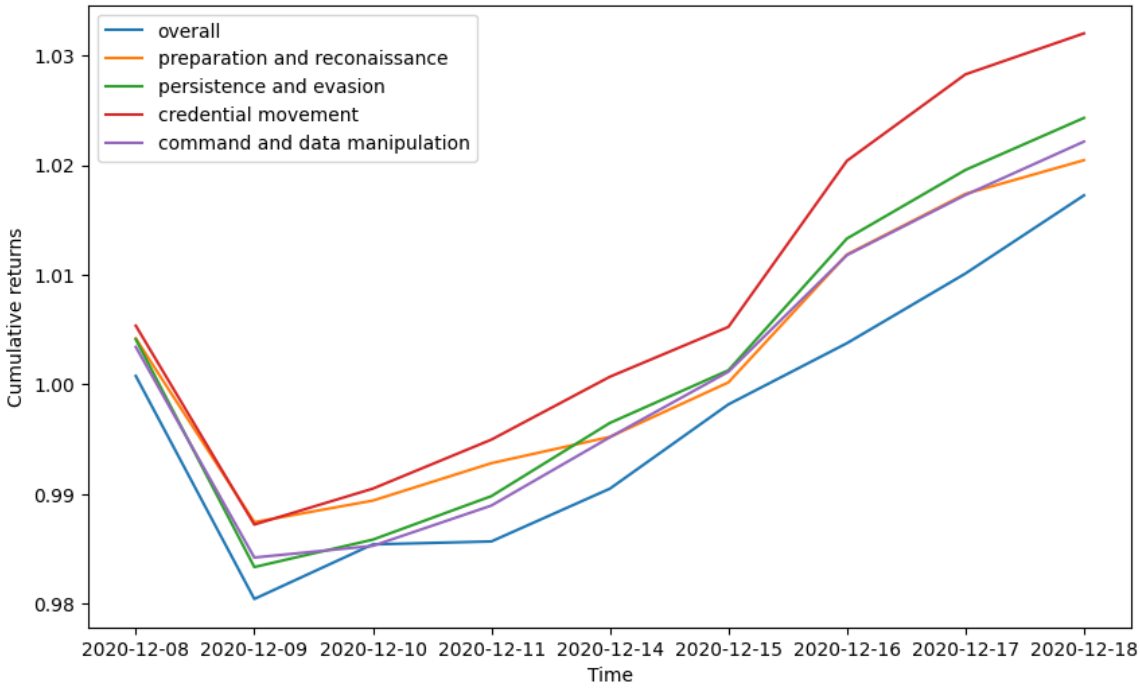

Fig. 23: **Cumulative returns of cyber-based portfolio (P20) around SolarWinds breach**

	CAR[-1,1]	CAR[-1,3]
overall	0.206	0.194
	[0.616]	[0.748]
preparation and reconnaissance	0.182	0.181
	[0.639]	[0.818]
persistence and evasion	0.192	0.189
	[0.533]	[0.675]
credential movement	0.198	0.200
	[0.565]	[0.735]
command and data manipulation	0.009	0.081
	[0.029]	[0.336]

Table 33: **Cumulative abnormal returns of cyber-based P5**

To estimate the cumulative abnormal returns (CAR), I use the market model around December 14, 2020, as $t = 0$. The beta of the market model is set up thanks to the returns of the prior year. The abnormal returns are given in percent. The t-statistics associated with the abnormal returns are given in the parenthesis.

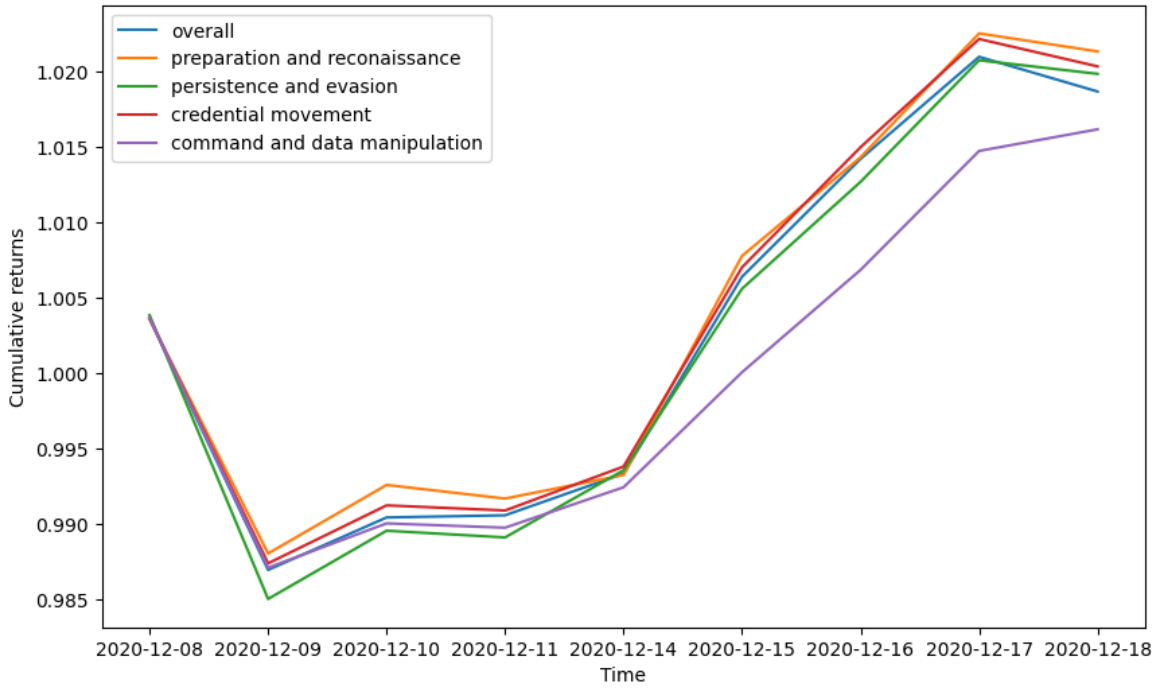

Fig. 24: **Cumulative returns of cyber-based portfolio (P5) around SolarWinds breach**

Appendix

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
Secrets	As defined in Florackis et al. (2023)	EDGAR - SEC
Volume per market cap.	Monthly trading volume / Market capitalization	CRSP
Humans per market cap.	Monthly number of employees / Market capitalization	Compustat and 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.

Risk/Uncertainty dictionary : risk, jeopardize, riskiness, risks, unsettled, treacherous, uncertainty, unpredictability, oscillating, variable, dilemma, perilous, chance, skepticism, tentativeness, possibility, hesitancy, unreliability, pending, riskier, wariness, uncertainties, unresolved, vagueness, uncertain, unsure, dodgy, doubt, irregular, equivocation, prospect, jeopardy, indecisive, bet, suspicion, chancy, variability, risking, menace, exposed, peril, qualm, likelihood, hesitating, vacillating, threat, risked, gnarly, probability, unreliable, disquiet, unknown, unsafe, ambivalence, varying, hazy, imperil, unclear, apprehension, vacillation, unpredictable, unforeseeable, incalculable, speculative, halting, untrustworthy, fear, wager, equivocating, reservation, torn, diffident, hesitant, precarious, fickleness, gamble, undetermined, misgiving, risky, insecurity, changeability, instability, debatable, undependable, doubtful, undecided, incertitude, hazard, dicey, fitful, tricky, indecision, parlous, sticky, wavering, unconfident, dangerous, iffy, defenseless, tentative, faltering, unsureness, hazardous, endanger, fluctuant, queries, quandary, niggle, danger, insecure, diffidence, fluctuating, changeable, precariousness, unstable, riskiest, doubtfulness, vague, hairy, erratic, ambivalent, query, dubious (Hassan et al., 2019)

Cyber score	**Covariance** $\cdot 10^3$	**Correlation**
persistence	0.2777	0.1038
command and control	0.3146	0.1281
impact	0.2960	0.1079
initial access	0.2237	0.0766
resource development	0.1841	0.0629
collection	0.2035	0.0723
exfiltration	0.1473	0.0499
credential access	0.2515	0.0916
privilege escalation	0.3088	0.1286
execution	0.2704	0.1121
defense evasion	0.1809	0.0761
reconnaissance	0.2145	0.0768
lateral movement	0.1272	0.0488
discovery	0.1640	0.0681
preparation and reconnaissance	0.1958	0.0710
persistence and evasion	0.1914	0.0795
credential movement	0.2350	0.0867
command and data manipulation	0.2114	0.0756
overall	0.1799	0.0699
sentiment	-0.0408	-0.0095

Table A2: **Cyber scores correlation and covariance with idiosyncratic volatility**

Correlation and covariance of the different cyber scores with the idiosyncratic volatility of the firms they are associated with. The idiosyncratic volatility at a given time is computed as the root squared of $var(\epsilon_i) = var(r_i) - cov(r_i, r_m)^2/var(r_m)$ taken over the last five years, where r_i and r_m are the excess return of the firm and the market. The covariance is multiplied by 10^3 to improve readability.