ARTICLE

<https://doi.org/10.1057/s41599-022-01039-1>

OPEN

The CoRisk-Index: a data-mining approach to identify industry-specific risk perceptions related to Covid-19

Fabian Stephany ^{1,2,3,9}, Leonie Neuhäuser ⁴, Niklas Stoehr⁵, Philipp Darius⁶, Ole Teutloff^{3,7} & Fabian Braesemann ^{1,2,3,8,9}✉

The global spread of Covid-19 has caused major economic disruptions. Governments around the world provide considerable financial support to mitigate the economic downturn. However, effective policy responses require reliable data on the economic consequences of the corona pandemic. We propose the CoRisk-Index: a real-time economic indicator of corporate risk perceptions related to Covid-19. Using data mining, we analyse all reports from US companies filed since January 2020, representing more than a third of the US workforce. We construct two measures—the number of ‘corona’ words in each report and the average *text negativity* of the sentences mentioning corona in each industry—that are aggregated in the CoRisk-Index. The index correlates with U.S. unemployment rates across industries and with an established market volatility measure, and it preempts stock market losses of February 2020. Moreover, thanks to topic modelling and natural language processing techniques, the CoRisk data provides highly granular data on different dimensions of the crisis and the concerns of individual industries. The index presented here helps researchers and decision makers to measure risk perceptions of industries with regard to Covid-19, bridging the quantification gap between highly volatile stock market dynamics and long-term macro-economic figures. For immediate access to the data, we provide all findings and raw data on an interactive online dashboard.

¹Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK. ²Humboldt Institute for Internet and Society Berlin, 10117 Berlin, Germany.

³Datenwissenschaftliche Gesellschaft Berlin, 10117 Berlin, Germany. ⁴RWTH Aachen University, 52062 Aachen, Germany. ⁵ETH Zürich, 8092 Zürich, Switzerland. ⁶Hertie School Berlin, 10117 Berlin, Germany. ⁷Copenhagen Center for Social Data Science, 1353 Copenhagen, Denmark. ⁸Saïd Business School, University of Oxford, Oxford OX1 1HP, UK. ⁹These authors contributed equally: Fabian Stephany, Fabian Braesemann.

✉email: fabian.braesemann@sbs.ox.ac.uk

Introduction

The Covid-19 pandemic has caused the largest global economic disruption of the twenty-first century (Fernandes, 2020; Zhang et al., 2020; Ozili and Arun, 2020). Travel bans, supply chain failures, and store closures pose significant risks to entire industries. Repeated lockdowns and relaxations of measures to fight the spread of the virus have left the overall economy in a stage of unprecedented uncertainty. In an attempt to mitigate the general economic downturn, governments have been mobilising considerable financial support (Werner et al., 2020; Zacharakis, 2020). While immediate general support programmes helped to stabilise economies ad-hoc, targeted and continuous countermeasures require reliable and up-to-date information on economic trends. However, the fast changing dynamics of the Covid-19 pandemic pose a challenge for economic forecasting and targeted policy intervention.

Owing to a lack of empirical data with granular time resolution, investigations on the economic impact of Covid-19 have mainly explored two pathways. Traditional macroeconomic research relies on metrics based on past economic shocks, such as economic simulations (del Rio-Chanona et al., 2020; Ludvigson et al., 2020). Another stream, following the paradigm of computational social science (Lazer et al., 2009), explores alternative data sources, such as stock market prices and returns (Ramelli and Wagner, 2020; Keogh-Brown et al., 2010; Buetre et al., 2006; Davis et al., 2020), news articles (Baker et al., 2020), website content (Kinne et al., 2020), search queries (Goodell and Huynh, 2020) or trade and transportation statistics as indicators of economic activity (Cerdeiro et al., 2020; Deb et al., 2021).

However, both approaches come with limitations in a crisis situation shaped by a rapidly adjusting economic environment. Traditional economic indicators, like composite indices (Lewis et al., 2020), manager surveys or labour market statistics, may have a sizeable time lag and they provide only an aggregated picture of the crisis. Empirical data with a high time resolution, on the other hand, such as stock markets, search queries, and news articles include a lot of noise and are prone to herd behaviour (Fama, 1965, 1991). Additionally, they could be detached from the economic reality faced by businesses. Moreover, many metrics do not allow for a granular analysis of the economic repercussions by industry. Examining current stock market dynamics, for example, may reveal general economic trends, whereas inter-sector differences may be obscured by irrational market dynamics. Moreover, stock prices represent the risk perception by all market participants, but the perspective of a company's management may provide a more isolated indication for real business risk outlooks and sector differences (Fama, 1980; Hassan et al., 2019).

To bridge the described gap in existing approaches, we propose a novel data source that allows a highly sensitive and reactive analysis of business risk perceptions. We investigate institutional language in business reports (10-K reports) filed to the U.S. Securities and Exchange Commission (SEC) using data-mining methods. SEC filings represent financial statements of large and medium-sized companies including a risk assessment. These reports are imperative to comply with legal and insurance requirements and therefore contain reliable risk evaluations of future risks. Prior work has underlined the forward-looking nature of SEC filings (Kogan et al., 2009; Richman et al., 2019; Chouliaras, 2015). Firms and managers report potential risk factors in the SEC filings to safeguard themselves against potential accusations of negligence (Lopez-Lira, 2021). Additionally, there is empirical evidence, which suggests that risk analysis can benefit from information inputs based on text data such as the SEC filings (Bochkay and Joos, 2021). Thus, we evaluate the reports as highly risk sensitive and forward-looking text-based data source.

To underline the relevance of our contribution, we evaluate our index with regard to effectiveness and efficiency of measuring industry-specific risk perceptions, similar to evaluations in medical trials (Wu et al., 2020) or public spending (Mandl et al., 2008). Efficiency is the ability to accomplish something with a relatively manageable amount of resources, e.g., wasted time, money, and effort or competency in performance. Effectiveness is defined as the degree to which something is successful in producing a desired result. We argue that our CoRisk-Index demonstrates effectiveness, as it anticipates unemployment developments across all observed industry sectors. The CoRisk-Index also correlates with established risk indicators like the CBOE Volatility Index (VIX)¹. Furthermore, we argue that the CoRisk-Index is also an efficient metric, because it requires less resources than other comparably effective approaches. For example, manager surveys such as the German IFO index require much more time and money than the automated CoRisk-Index².

To our knowledge, only one other study (Hassan et al., 2020) attempts to quantify risk perceptions related to Covid-19 (and other recent virus pandemics) at the firm-level. The authors explore to what extent firms build resilience to pandemic-induced economic shocks on a global scale. The study builds on quarterly earnings conference call protocols. These documents contain risk perceptions brought up during the calls. However, Hassan et al. focus on the firm-level, while we put emphasis on the industry level. In addition, compared to the call protocols investigated by Hassan et al., the 10-K reports investigated here come with some advantages. 10-K reports represent legally binding reports including a dedicated risk section; they are publicly available allowing for full reproducibility; and they follow a standard format allowing for automated text mining.

In collecting all 10-K reports published since 30th January 2020—the day the term *coronavirus* first appeared in a 10-K report—we assess and track the reported risk perceptions related to Covid-19 for the different sectors³ of the US economy. The reports analysed here cover the risk assessments of U.S. companies representing more than 44 million U.S. employees, i.e., one-third of the total U.S. economy. We construct two measures: (a) the number of 'corona' words in each 10-K report, and (b) the average *text negativity* of the sentences mentioning corona in each industry. The text negativity is calculated as the share of negative words (Loughran and McDonald, 2011) per corona-mentioning sentence. The geometric mean of corona-words and text negativity builds the CoRisk-Index, which is calculated for each industry. Moreover, to understand how companies in different industries assess Covid-19 as a risk factor over time, we use natural language processing to identify the topical context in which Covid-19 related risks factors are described (for details on the topic selection, see Methods section 'Topic detection').

As we will demonstrate, the CoRisk-Index represents a highly informative, reliable, up-to-date measure and open data source. The data can help researchers and decision makers to understand the economic repercussions related to Covid-19 in different industries. It closes a data gap between short-term financial market indicators and macroeconomic statistics in order to gain a more complete understanding of the different aspects of the economic crisis caused by the corona pandemic and to design better targeted responses. For immediate access to all raw data we offer an interactive online dashboard.⁴

This paper illustrates how to transform data from a publicly accessible source into an economically meaningful measure of industry-specific business risk perceptions related to the Covid-19 pandemic: the CoRisk-Index. Our work does not aim to predict future developments, but to measure previously unobserved risk perceptions of industries with regard to Covid-19. Besides the

Fig. 1 The CoRisk measures reflect economic concerns with regards to Covid-19 since early 2020. **A** Daily cases of Covid-19 infections in the United States (source: www.ourworldindata.org): the first wave of the pandemic started in March 2020, the second in September. **B** S&P 1200 Global Index (source: www.spglobal.com): As the Covid-19 pandemic reaches the level of a global health crisis, stock markets plummet on 20th February but recover sharply from April 2020 onward. **C** Number of 'corona' keywords in 10-K reports: the number of reports about the virus starts to rise before the first wave and the measure shows oscillating behaviour with peaks in June, September and December 2020. **D** Share of negative words in 'corona' sentences of 10-K reports: The negative sentiment surges before the global stock market crash and peaks before 'Black Monday's' most severe stock market losses.

forward-looking nature and risk-sensitivity, the text data promises additional signals that may complement macroeconomic models to forecast and measure economic crises. In order to test the efficiency and effectiveness of the CoRisk-Index as an indicator of economic risk perceptions we conduct Granger causality tests based on vector autoregression models with US unemployment statistics, the S&P 1200 Global Index and the VIX volatility indicator⁵. The results from the Granger causality tests confirm the effectiveness of the CoRisk-Index in describing economic risk perceptions related to Covid-19⁶.

The remaining paper proceeds as follows. In the next section, we introduce the data and CoRisk-Index in more detail. Subsequently, we discuss implications and critically reflect upon the limitations of our approach. The Methods section provides a comprehensive step-by-step explanation of the data collection, preprocessing, and analysis.

Results

'Corona' keywords and text negativity: early warning signals of the economic shock. The 10-K reports filed to the SEC contain early warning signals of the economic shock that occurred throughout the Covid-19 pandemic. The risk perception captured by the reports shows early reactions to the unfolding crisis, even before the number of Covid-19 cases surged in the United States (Fig. 1A). Before the first wave hit the country in March 2020, the number of 'corona' keywords in 10-K reports had been rising steadily (Fig. 1C). Subsequently, the number of corona keywords shows an oscillating pattern with peaks in June, September, and December 2020, reflecting waves of economic concerns about the pandemic, which overlap roughly with the infection patterns in 2020.

Similarly, before the stock markets plummeted on February 20th (Fig. 1B), the text negativity (average share of negative words in sentences mentioning 'corona' or 'covid' in the 10-K reports) spiked in early February (Fig. 1D). It peaks at a share of 7%

negative words per corona-sentence just before the most severe stock market losses ('Black Monday') on March 9th. The text negativity quantifies the sentiment with which firms report about Covid-19 related business issues. The sharp increase in the first quarter of 2020 reflects the uncertainty in global economic outlooks, when businesses suddenly became aware of the pandemic turning from a regional into a global health crisis. In the subsequent months, as economic aid packages and financial support programmes started to back the capital markets, stock markets started to grow continuously, which is mirrored by a steady decrease of the text negativity of corona-related sentences.

Constructing the CoRisk-Index. Based on the count of 'corona' keywords and the text negativity in 10-K reports, we construct the *CoRisk-Index* (Fig. 2A) for eight out of 13 sectors of the U.S. economy, which are well represented by SEC filings (for details on the industry classification, 10-K reporting patterns, the calculation of the index, and the selection of industries, see the Methods section from 'Matching industry classification systems' to 'Representativeness'). The CoRisk-Index is calculated as the geometric mean of the total count of 'corona' keywords (see Fig. 1C) and the text negativity (see Fig. 1D) per industry. The index serves as a granular economic indicator measuring industry risk perceptions related to Covid-19. The index shows how sectors differ in their timing and magnitude of risk awareness. For example, the indices of the sectors Manufacturing, Wholesale & Retail, and Professional & Business Services (first row in Fig. 2A) show a steep and early increase in the first two quarters of 2020 up to a plateau on a relatively high level between 100 and 150 index points on which they stayed thereafter. In contrast, the industries Leisure & Hospitality, Finance, and Transportation & Utilities (second row in Fig. 2A) started to rise later, but with a steady growth throughout 2020. Two other industries, Information and Mining (third row), have peaked on relatively low levels and moved between 50 to 100 index points since then. The 'All

A The CoRisk-Index: a composite index measuring the industry-specific business risks related to Covid-19

B CoRisk topics: Share of topic-specific keywords in 'corona' sentences

Fig. 2 The CoRisk-Index reveals the dynamics of the economic crisis in different industries. A The CoRisk-Index: a composite measure of industry-specific business risks. It shows how industries differ in timing and magnitude of risk awareness towards Covid-19 together with smoothed trend lines (dotted). **B** CoRisk topics: the share of domain-specific keywords in 'corona' sentences of 10-K reports. The data reflects the relevance of the different dimensions of the crisis for the eight considered industries over time.

Industries' panel summarises the behaviour of the individual industries and reveals three peaks in early Summer, in Autumn, and Winter 2020.

Applying automated text analysis to the sentences in the 10-K reports that mention 'corona' allows us to investigate how the pandemic influenced the risk perceptions of the different industries in particular contexts (Fig. 2B).⁷ For example, as the U.S. government posed first travel restrictions on 31st January 2020, travel-related reporting in Transportation & Utilities surged, while other industries, such as Professional & Business Services, were less concerned about travel restrictions. Similarly, the subsequent supply and production disruptions due to the restrictions in East Asia led to a sharp rise of concerns in the Manufacturing industry. They did not substantially concern more service-oriented sectors such as Leisure & Hospitality or Information. The demand shock due to the first lockdown, on the other hand, led to severe worries in the Wholesale & Retail industry, not so in less-consumer focused sectors such as Mining.

The potential financial impacts of the crisis were also differently perceived by the industries. The Finance sector showed an early rise of concerns in the first quarter of 2020, but remained relatively stable throughout the rest of the year. Only in January 2021, the Finance sector started to focus more on the financial impacts of the pandemic. In contrast, the Mining sector has started to report substantially more about the financial repercussions of the crisis since the start of the second wave of the pandemic in Autumn 2020, potentially reflecting the clouded prospects of the world economy. Following the 'Coronavirus Aid, Relief and Economic Security (CARES) Act' from March 2020, which released more than USD 500 billions in loans to the U.S. economy, the Wholesale & Retail sector has started to mention aid-related keywords in their reports, with more reporting in times of peaking infection rates in late spring and late Autumn 2020. The Leisure & Hospitality industry, which could not recover substantially during the short period of decreasing infection rates in Summer 2020, intensified the reporting of aid-related keywords throughout the second half of 2020. The fact that remote working schedules are only relevant in those sectors without face-to-face interactions is reflected in the keyword pattern of the Professional & Business Services industry with regards to keywords around remote working. These surged to high levels in the periods of stay-at-home policies and decreased, for example, during Summer 2020, when the infection rates were lower. In comparison, other industries, such as Transport & Utilities, have not been concerned with remote working topics to such an extent, possibly because teleworking is, due to the nature of work carried out, not that relevant in certain sectors. The overall rising importance of the vaccine to overcome the crisis is reflected by the rising trend of vaccine-related keywords in all industries.

Effectiveness and generalisability of the CoRisk-Index. The CoRisk-Index fills a data gap between short-term financial information, as captured by stock markets, and long-term macroeconomic data. While macroeconomic statistics, such as GDP, production, and cost information of individual sectors are able to describe the full spectrum of the economic crisis, they are available only with a substantial time lag and with a relatively low resolution. Many industry-level statistics are reported only quarterly. Drastic changes in the business outlooks in shorter time frames, as they have been common during the pandemic, can therefore not be observed. In contrast, with its high time resolution, the CoRisk-Index shows fluctuations and oscillating patterns that reflect the dynamically evolving economic perspectives during the ups and downs of the crisis. At the same

time, the index provides information that is related to industry-specific macroeconomic indicators, i.e., unemployment rates. In the following, we show that the CoRisk-Index is an effective measure of economic risk perceptions. Moreover, we illustrate that the methodology can be generalised to other contexts such as the 'US-China trade war' of 2018.

As we have illustrated in Figs. 1 and 2, the CoRisk-Index effectively measures economic risk perceptions related to Covid-19. As an automated text-mining approach is most likely more efficient in measuring risk perceptions than repeated manager surveys, we need to compare the CoRisk-Index to other similarly efficient metrics, such as stock market performance or volatility indices, such as the VIX. To formally test the forward-looking nature of the CoRisk-Index, in comparison to these metrics, we perform a VAR-based Granger causality test for explaining general and sector-specific unemployment. As described in greater detail in the Methods section 'Granger causality test', the null hypothesis of the Granger causality test assumes that the lagged values of an additional time series do not add any explanatory power to an auto-correlation model alone. Figure 3A summarises the results of the Granger causality tests. First, the KPSS stationarity test confirms that, after normalisation, all times series are stationary—a basic requirement for Granger causality testing. The Granger causality test statistics (more details in Tables 1 and 2) show that the CoRisk-Index reveals Granger causality with the VIX, the S&P 1200 Global Index and, in particular, changes in unemployment rates, both on the general economic and industry-specific level. The CoRisk-Index and its respective industry-specific sub-indices lead unemployment developments by between 1 and 4 months⁸.

The methodology underlying the CoRisk-Index can be generalised beyond the case presented here. Reports filed to the U.S. Securities and Exchange Commission represent a rich source of textual data reflecting political and economic events, which can be used to track the influence of external events on the business environment. We exemplify the generalisability of the approach by looking at the U.S.-China trade war of 2018. Figure 3B shows the text negativity of 'china' sentences, i.e., share of negative words in sentences mentioning the keyword 'china'. The negativity varies between 1.7% and 3.4% over the course of the trade war, with tariff announcements being reflected by periods of sharply increasing negativity and relaxing events such as trade talks manifesting themselves as decreasing text negativity. The example illustrates that the CoRisk methodology provides a flexible tool that is able to track business risk perception in various contexts. See the Methods section 'Historical robustness' for historical robustness checks in which we compare the text negativity of 10-K reports and macroeconomic variables between 2000 and 2018.

Discussion

Governments aim to counterbalance the global economic crisis induced by the Covid-19 pandemic with cyclical and fiscal policy packages of enormous volumes. The CoRisk-Index represents an attempt to contribute to the pressing demand for empirical data on the economic impact of the ongoing pandemic. It equips policy-makers and researchers with a tool to analyse and track industry-specific risk perceptions related to Covid-19, based on data that represents a large part of the U.S. economy. Our analysis shows that industries vary greatly in their risk perception as well as with regards to the topics that they are concerned with. This underscores the necessity to incorporate alternative data sources of up-to-date empirical information to better understand the differential effects of Covid-19 on individual sectors of the economy. Thereby CoRisk data helps to assess the economic

A Results of Granger causality tests assessing the correlation between the CoRisk-Index and VIX, S&P 1200 Global Index and monthly unemployment rates

Independent Variable	Dependent Variable	KPSS stationarity test, $\alpha = 5\%$	Statistically significant time lag, according to GCT	Granger Causality F-Test, $\alpha = 5\%$
CoRisk-Index	VIX	✓	1	✓
CoRisk-Index	SuP 1200	✓	1	✓
CoRisk-Index	Unempl. (All Industries)	✓	2	✓
CoRisk (Finance)	Unempl. (Finance)	✓	1	✓
CoRisk (Information)	Unempl. (Information)	✓	1	✓
CoRisk (Leis. & Hosp.)	Unempl. (Leis. & Hosp.)	✓	4	✓
CoRisk (Manufacturing)	Unempl. (Manufacturing)	✓	2	✓
CoRisk (Mining)	Unempl. (Mining)	✓	0	
CoRisk (Prof. Serv.)	Unempl. (Prof. Serv.)	✓	1	✓
CoRisk (Transport.)	Unempl. (Transport.)	✓	1	✓
CoRisk (Retail)	Unempl. (Retail)	✓	1	✓

B Generalisation of the CoRisk methodology: share of negative words in 'china' sentences in 2018 SEC filings

Fig. 3 The CoRisk data reflect macroeconomic developments. A Results of Granger causality tests assessing the correlation between the CoRisk-Index and VIX, S&P 1200 Global Index and monthly unemployment rates. We find that the marginals of all-time series are stationarity based on a KPSS test. The Granger causality tests (detailed statistics and the monthly aggregated data used in the tests in Tables 1 and 2) yield significant results between the CoRisk-Index and unemployment as well as VIX and S&P 1200. This means that values of the CoRisk-Index are correlated with unemployment values, VIX and S&P 1200 during the course of the pandemic. **B** Generalisability of the CoRisk methodology: share of negative words in 'China' sentences in 10-K reports during the 'US-China trade war' in 2018 (solid line and grey dotted trend line). The methodology allows to extract the china-related negative sentiment from the reports and it reflects negative (red) and positive (green) events during the crisis, pointing towards the general applicability of the tool to reveal granular insights about the effects of events on businesses.

effects of the crisis and to design viable support instruments targeting the most affected industries.

For example, the first phase of the economic crisis in spring 2020 consisted of several related shocks that affected individual parts of the economy in different ways. Travel restrictions, the first sign of the crisis, were more relevant for the Transportation & Utilities sector than for more domestically oriented industries. The travel restrictions were soon followed by supply and production disruptions with most severe impacts to the globally connected value chains of the Manufacturing industry. The store closures and curfews of the first lockdown in March, however, resulted in a severe demand shock that was more concerning for Wholesale & Retail than for those industries that could largely continue their business operations.

As the crisis continued throughout 2020, with governmental reactions, restrictions, and relaxations alternating, the different industries were again confronted with particular challenges. For instance, while the direct financial impacts of the lockdown measures were reduced due to large-scale economic aid packages, concerns about the short- and long-term economic consequences started to rise in different sectors. While Wholesale & Retail saw their business situation, and hence their need for governmental support, relaxing in Summer 2020, companies from Leisure & Hospitality continued to feel increased business stress throughout the year. The lockdown-induced trend towards remote working, on the other hand, did play a significant role in office-based sectors such as Professional & Business Services.

Table 1 Aggregated monthly data used in Granger causality test.

Date (Year-month)	CoRisk Index	Unemployment Rate (%)	VIX Index	S&P 1200 Global Index
2020-01	0	4	14	2655
2020-02	21	3.8	20	2633
2020-03	62	4.5	58	2116
2020-04	95	14.4	41	2175
2020-05	109	13	31	2282
2020-06	119	11.2	31	2451
2020-07	105	10.5	27	2528
2020-08	114	8.5	23	2649
2020-09	107	7.7	28	2632
2020-10	103	6.6	29	2655
2020-11	130	6.4	25	2787
2020-12	105	6.5	22	2924
2021-01	81	6.8	25	3021
2021-02	155	6.6	23	3085
2021-03	145	6.2	22	3090
2021-04	84	5.7	17	3230
2021-05	85	5.5	20	3260
2021-06	75	6.1	17	3319
2021-07	18	5.7	18	3354
2021-08	0	5.3	17	3404

Implications and use cases. These examples underline the value of the granular and topic-specific business risk information made accessible via the CoRisk-Index. The index fills a data gap between the aggregated financial signals from the stock market and the backward-looking macroeconomic indicators provided in official economic statistics. The stock market signal is biased towards short-term fluctuations and the signal can be detached from the overall economic situation. For example, the stock market's rally to all-time highs since April 2020 does not reflect the overall uncertainty many businesses are still facing. Macroeconomic indicators, such as unemployment rates or aggregated industry revenues, on the other hand, might not reflect the actual situation either. Short-term fiscal policies, like furlough schemes or governmental investment programmes, might conceal the full extent of the crisis.

The CoRisk-Index can help to identify the particular challenges of different industries and guide decision makers' attention to those risk factors. For example, the CoRisk data could have helped to reveal supply-related problems in the Manufacturing industry early on, which might have helped to design policies to reallocate the production of relevant parts and raw materials to domestic sites. The data would have allowed to identify those industries most in need of support for the set-up of IT infrastructure to enable remote working, as one measure to reduce unnecessary face-to-face interaction in sectors such as Professional & Business Services. Moreover, the CoRisk-Index would have also helped to differentiate the timing of immediate economic aid packages. For instance, while the Wholesale & Retail industry reported severe concerns particularly in the first weeks of the lockdown, other sectors such as Mining, or Leisure & Hospitality expressed more need for support in Autumn 2020; an indication that could have been used to define the sequence of economic support packages by industry risk groups, similar to the prioritisation of risk groups in the vaccination programme. Throughout the further development of the crisis, the CoRisk data will cover upcoming issues and identify those parts of the economy that are most affected by these issues.

Limitations and outlook. We limit our investigation to 10-K reports as a source of company risk reporting. Kogan et al. (2009) While the decision to use 10-K reports comes with the advantages of standardised and comparable reporting, as well as with full public access to the data, it limits the scope of the analysis to the United States. Unfortunately, comparable international databases that would contain similar reporting as the 10-K filings are not publicly available, and other business reports, such as quarterly earnings conference calls, might not adhere to the same strict reporting requirements as the 10-K reports. Too much variation in the reporting structure and style would limit the value of the context-specific topic analysis, which we perceive as a key contribution of our study. In other words, while the scope of the work had to be restricted to the US economy and one annual report type, we are confident that this restriction is outweighed by the standardised nature of the public data. This allows for better interpretation and full reproducibility of the results that are made available in open source format. Lastly, our approach relies on company self-reporting and risk disclosure, which may affect a company's stock price volatility (Wang and Xing, 2020; Theile et al., 2020) The long-term implications of the Covid-19 crisis are still uncertain. Our approach approximates economic repercussions based on the assessment by company experts in the different sectors. Naturally, this does not include risks that are unforeseeable for companies themselves at the time of reporting.

Nonetheless of its limitations, we believe that the CoRisk-Index represent a valuable source of empirical information about the issues faced by various industries during different stages of the pandemic. The risk of a global health crisis calls for a constant re-assessment of economic repercussions and the corresponding policy responses. In this situation, agile computational social science methods such as text mining and natural language processing allow us to extract valuable information at low-cost from highly sensitive text-sources such as the 10-K SEC reports. As the crisis unfolded, we continuously extended our work. The CoRisk-Index is updated weekly until the end of the pandemic, in order to support fellow researchers and policy-makers with timely data on industry-specific risk perceptions. In addition to the findings presented in this paper, we provide detailed industry-specific insights and all raw data via an interactive online dashboard.⁹

Methods

Data collection

Automated extraction of 10-K reports. The U.S. Securities and Exchange Commission (SEC) stores all reporting in a central repository¹⁰. Here, users can access meta-level information, such as index files, e.g., lists of all reports issued in the second quarter of 2020 directly. Alternatively, individual filings and meta-level information can be retrieved via various statistical packages, e.g., 'edgar' (R) or 'sec-edgar-downloader' (python), or a freemium API¹¹.

However, several limitations make the use of these ready-made devices impractical for the specific research purposes of this work. Downloading all relevant reports as .txt/.html files via the ready-made packages is possible in theory but requires a lot of time and hard disk space. In addition, the .txt file contains several unwanted css/html code patterns that make the identification of corona-sentences and the counting of words from them unreliable.

Alternatively, we initially refer to the *crawler.idx* index file¹² in the SEC repository. The index file holds a full list of all reports issued in a given quarter. This document shows the meta-level *index.htm* page for each company. On this page, the most recent

Table 2 Granger causality test results.

Indep.	Dep. Variable	Lags	ssr-based <i>F</i> - test <i>F</i> -val.	ssr-based χ^2 -test χ^2 -val.	Likelihood ratio test χ^2 -val.	Parameter <i>F</i> - test <i>F</i> -val.
CoRisk	VIX	1	213***	255***	49.0***	213***
		2	0.57	1.62	1.55	0.57
		3	0.66	3.51	3.18	0.66
		4	0.21	2.05	1.92	0.21
		5	0.53	12.46	8.91	0.53
CoRisk	S&P 1200	1	93.6***	112***	35.6***	93.6***
		2	0.22	0.62	0.61	0.22
		3	0.91	4.84	4.23	0.91
		4	0.44	4.39	3.85	0.44
		5	0.40	9.22	7.08	0.40
All industries CoRisk	Unempl.	1	0.07	0.08	0.08	0.07
		2	359***	1019***	69.9***	359***
		3	0.66	3.52	3.18	0.66
		4	0.23	2.29	2.13	0.23
		5	2.64	61.5***	23.6***	2.64
Finance CoRisk	Unempl.	1	15.9**	19.4***	12.9***	15.9**
		2	3.97	11.5***	8.69**	3.97
		3	2.94	16.5***	11.2***	2.94
		4	2.19	24.5***	14.2***	2.19
Information CoRisk	Unempl.	1	38.5***	48.2***	21.6***	38.5***
		2	0.01	0.02	0.02	0.01
		3	0.39	2.54	2.32	0.39
		4	0.25	3.97	3.43	0.25
Leisure and hospitality CoRisk	Unempl.	1	0.35	0.44	0.43	0.35
		2	1.83	5.69	4.78	1.83
		3	3.61	23.5***	13.4***	3.61
		4	11.6**	183***	33***	11.5**
Manufacturing CoRisk	Unempl.	1	0.03	0.04	0.04	0.03
		2	130***	380***	51***	130***
		3	0.16	0.89	0.87	0.16
		4	1.5	16.4	10.8**	1.46
Mining CoRisk	Unempl.	1	0.56	0.69	0.68	0.56
		2	0.13	0.39	0.38	0.13
		3	0.17	1.04	1.00	0.17
		4	0.39	5.14	4.33	0.39
Professional and business services CoRisk	Unempl.	1	50.9***	62.7***	25.5***	50.9***
		2	0.53	1.58	1.5	0.53
		3	0.07	0.41	0.40	0.07
		4	0.37	4.79	4.08	0.37
Transportation and utilities CoRisk	Unempl.	1	18.9***	23.6***	14.2***	18.9***
		2	0.47	1.45	1.38	0.47
		3	0.26	1.69	1.59	0.26
		4	1.01	16.2***	10.2**	1.01
Wholesale and retail CoRisk	Unempl.	1	58.8***	73.6***	26.6***	58.8***
		2	0.43	1.32	1.27	0.43
		3	0.65	4.25	3.68	0.65
		4	0.67	10.6**	7.6	0.66

****p* < 0.01, ***p* < 0.05

10-K report is linked¹³. Unfortunately, the htm-version of each 10-K report has a cryptic file name that is not associated with the company, date or industry parameters. Hence the currently implemented scraper (*python 2.7*) pipeline of our project, first, fetches the list of recently listed 10-K reports from *crawler.idx*,

secondly, constructs the meta-level index.htm for each company and finds the link to the most recent 10-K report. The algorithm then scrapes the report text, before lastly, identifying the sentences related to corona (examples of text element that contain the term 'coronavirus' are displayed in Fig. 4).

CBRE

Adverse economic conditions or political or regulatory uncertainty or significant public health events, such as pandemics, could also lead to a decline in leasing volume, property sales prices, funds invested in commercial real estate assets or planned development activity, which in turn could reduce the commissions and fees we earn. During 2019, our Asia Pacific business experienced declines in leasing activity amid rising geopolitical and trade uncertainty and slowing regional economies. Furthermore, in December 2019, a strain of coronavirus was reported to have surfaced in Wuhan, China, resulting in decreased economic activity in China and concerns about a potential pandemic, which would adversely affect the broader global economy. At this point, the extent to which this coronavirus may impact the global economy and our results is uncertain, but pandemics or other significant public health events, or the perception that such events may occur, could have a material adverse effect on our business.

BBX Capital

Risks that public health issues, such as the recent coronavirus outbreak, and natural disasters including hurricanes, may adversely impact the Company's financial condition and operating results, including, with respect to the recent coronavirus outbreak, impacts resulting from disruptions in the Company's supply chains, potential declines in leisure travel and consumer traffic at malls and retail locations, and public health concerns related to bulk candy products;

CORONA
TRADE
SUPPLY
TRAVEL

Fig. 4 Examples of paragraphs in 10-K reports that mention the term 'coronavirus' and other business-specific terms.

All sentences and their respective report properties are stored for later processing.

Specification of corona keywords. Since this study examines the attention attributed to COVID-19 in the SEC filings, the discovery mechanism of relevant COVID-19 mentions is of central importance. To mitigate susceptibility to errors due to word splitting, stemming and other text preprocessing, we decided for the most simple approach based on the matching of regular expressions. We scanned the reports for the two relatively unambiguous terms 'corona' and 'covid', also accounting for 'coronavirus' and 'covid-19' without duplication. For this process, the entire text is set to lower case.

Topic detection. We apply unsupervised methods for topic detection from Natural Language Processing to further analyse the reports with regard to topic related risks. Different sectors are facing different challenges, therefore companies are reporting about different corona-related risks. We aim to capture these risk topics via a keyword search on predefined topics. In order to explore possible topics, we used Latent Dirichlet Allocation (LDA) for unsupervised topic modelling, similar to Dyer et al. (2017). We only apply the topic model to corona-related paragraphs in the risk sections. We additionally examine the most frequent words and bi-grams in the documents. Using this exploratory analysis, we define a set of topics, which are specified by keywords. We then conduct a keyword search to count how much these terms are mentioned in the different industries in order to estimate the topic prevalence.

Unsupervised topic modelling. We use unsupervised learning techniques to explore the space of topics that companies discuss when describing coronavirus-related risks. Latent Dirichlet Allocation (LDA) is a Bayesian computational linguistic technique that identifies the latent topics in a corpus of documents (Blei, 2012). This statistical model falls into the category of generative probabilistic modelling: a generative process, which defines a joint probability distribution over the observed random variable, i.e., the words of the documents, and the hidden random variables, i.e., the topic structure. In other words, LDA uses the probability of words that co-occur within documents to identify sets of topics and their associated words (Dyer et al., 2017). The number of topics has to be defined in advance. LDA is a frequently used technique to identify main topics in a corpus. Nevertheless, the interpretation of these topics can sometimes be difficult. We thus perform LDA for explorative purposes in our research only and apply the following steps:

Sample restriction: We filter all sentences from the risk sections that mention either 'corona' and 'covid', thereby also accounting for 'coronavirus' and 'covid-19'.

Text preparation: Before we train the LDA model we prepare the documents to achieve better performance of the method. We remove all common English stopwords, which are frequent words such as 'is', 'the', and 'and' as well as those words, which appear in at least 80% of the documents. These words are not useful in classifying topics as they are too frequent and therefore decrease performance. Moreover, we delete all words that do not occur in at least two documents.

LDA: We turn the documents into numerical 'Bag of words' feature vectors, disregarding word order. We then use LDA to extract the topic structure. Like any unsupervised topic model, this requires setting the number of topics a priori. We selected this key parameter based on semantic coherence, evaluating a range of two to eight topics leading to a final model of four topics. The top ten terms of each topic are displayed in Table 3.

Dictionary-based topic search. The topics that are derived from the unsupervised topic modelling give a good insight into the general narratives of risks used in the documents. Nevertheless, they are hard to interpret, as early corona-related risk reports are still generic in that various risk factors are covered. Topic four, for example, provides an unspecified context with regard to the outbreak of the illness in China. Similarly, topic three covers the potential impact of the crisis in an unspecified context. In contrast, most of the business and economics related keywords appear to be covered in topic two. Moreover, the unsupervised methods are not deterministic. To ensure a robust and comparable topic identification over time, we use a dictionary-based topic search for the final statistics. A list of final topics with defining keywords is displayed in Table 4 and derived as follows: We identify the most frequently used bi- and trigrams tokens across all report documents. We then group the most frequently occurring, non-trivial tokens—excluding paraphrases and such—into topic areas. The organisation of topics and attribution of keywords is informed by both the previous exploratory unsupervised topic modelling and additional expert knowledge from economics. This two-step machine-human decision making process reduces the risk of arbitrary topic selection, while still enabling contextual refinements based on relevant (human) domain knowledge. Using the resulting keyword lists for the topics, we can conduct a dictionary-based keyword search in the filings. We measure the topical context of each corona-sentence by calculating the share of topic keywords relative to the word length of the sentence. This metric is later aggregated for industries and specific points in time.

Matching industry classification systems. The SEC classifies firms into industries using an amended version of the 1987 Standard Industrial Classification (SIC).¹⁴ The SIC is a system for classifying industries by a four-digit code. It was replaced by the

Table 3 Results of an LDA topic modelling with four topics. The algorithmically derived topics give a good insight into the general narratives of risks used in the documents.

Topic number	Top 10 words
Topic 1:	Impact, extent, including, outbreak, uncertain, future, highly, results, developments, depend
Topic 2:	Operations, including, health, outbreak, business, supply, products, economic, public, result
Topic 3:	Outbreak, spread, countries, impact, including, china, business, potential, economic, government
Topic 4:	China, outbreak, novel, covid, adversely, wuhan, strain, business, december, recent

Table 4 Keywords for eight topical domains. We combine the results of the unsupervised methods with domain knowledge from economics to label the eight main topics and specify defining keywords by selecting from the most frequent bi-and trigrams.

Topic	Keywords
Production	Business operation, business disruption, product, work stoppage, labour disruption, labor, work, manufacturing operation, labor shortage, employee productivity, product development, business activity
Supply	Manufacturing facility, manufacture facility, contract manufacturer, service provider, logistic provider, supply disruption, party manufacturer, supply disruption, facility, supply, transportation delay, delivery delay, supplier, business partner, supply chain, material shortage
Travel	Air travel, travel, travel restriction, airline industry, travel disruption
Demand	Store closure, distribution channel, market condition, consumer spend, market acceptance, consumer confidence, consumer demand, consume, store, customer, store traffic
Finance	Operating result, cash flow, stock price, estate value, credit availability, performance problem
Aid	Aid, coronavirus aid, care act, cares act, paycheck protection
Remote	Distancing, stay-at-home, remote
Vaccine	Vaccin, recovery, treatment, reopen

North American Industry Classification System (NAICS) in 1997. The NAICS system is still in use, in its last revision from 2017, and it is being used by US government agencies. Thus, it is important to match the SIC codes used by the SEC with the NAICS system, in order to allow researcher and economists to use the CoRisk data in industry-specific applications.

To match the different classification systems, we use merging tables provided by the US Census Bureau.¹⁵ The code (R) and the merging tables are available on GitHub.¹⁶ After having merged the four-digit SIC codes, we use the NAICS overview provided by the US Census Bureau to identify the larger two-digit sectors, on which the CoRisk data are aggregated.¹⁷ During merging, some industries are assigned to several categories. In order to drop duplicates, we only keep the most frequent section per SIC code. While it might be possible that some information is lost during the merging process from the 1987 four-digits SIC system to 2017 two-digit NAICS sectors, we assume that less categories, which are compatible with the system used in current US statistics, provide a better overview of relevant processes than an aggregation, which is too fine-grained and which consists of low sample sizes within each category.

10-K reporting pattern

Filing of 10-K reports to the SEC. Companies with more than 10 million USD in assets or a class of equity securities that is held by more than 2000 owners must file annual 10-K reports to the SEC, regardless of whether the securities are publicly or privately traded¹⁸. All 10-K reports are made publicly available by the SEC. In particular, but not exclusively, in the risk section of the report, the company outlines potential threats to its business such as likely external effects, possible future failures to meet obligations, and any other risks disclosed to adequately warn investors and potential investors. Companies are required to use 'plain English' in describing these risk factors, avoiding overly technical jargon that would be difficult for a layperson to follow.

Comparing 10-K and 10-Q reports. In addition to annual 10-K reports, the SEC requires companies to publicly disclose their actions in a set of reporting formats (10-X). Apart from historical reporting standards and niche reporting categories, company disclosures with the SEC can be separated in two groups: Annual 10-K and quarterly 10-Q reporting. One of the four quarterly 10-Q reports is subsumed by the annual 10-K report.

We limit our analysis to the information contained in 10-K reporting for the following reason. On March 4, the SEC has explicitly advised public companies to assess what the coronavirus means for their future operations and financial results and to make appropriate disclosures to their shareholders and other members of the investment community. Furthermore, the SEC encouraged companies to delay SEC filings if necessary to develop the information required to make accurate and complete disclosures of the impact of the coronavirus on its operations and financial conditions. Specifically, the SEC issued an order stating that public companies that are unable, because of the coronavirus, to meet filing deadlines for SEC reports due to be filed March 1 to April 30, 2020, will have 45 additional days to file these reports so long as, among other things, they file reports on Form 8-K describing the reasons why the report may not be filed on a timely basis¹⁹. This announcement has caused and interference with the normal reporting procedures in at least two ways. Companies delayed their quarterly reporting until the last possible date (April 30, 2020) and companies were incentivised to 'talk' about Covid-19 related issues.

From our perspective, this reporting bias should manifest most strongly in 10-Q reports, as they are more susceptible to short-term changes in reporting standards. We decide to include solemnly 10-K reports in our CoRisk analysis, as they have a long-term outlook of one year and a lower susceptibility to ad-hoc changes in reporting standards compared to 10-Q reports (see the Figs. 5 and 6).

Elements of the CoRisk-Index. The CoRisk-Index reflects the risk perception of industry sectors towards Covid-19. The measure

Fig. 5 The CoRisk-Index based on 10-K reports (left panel) and on 10-K and 10-Q reports (right panel). The indices show some similarity, however the index on the right is biased by the very high number of 10-Q reports being filed in May 2020.

Fig. 6 Number of reports filed to the SEC per day January to May 2020. The 10-Q reports show a very different dynamic from the 10-K reports. While 10-K reports are filed throughout the first quarter of 2020 (and to a lower extent in the second quarter), the number of filed 10-Q reports is vastly skewed towards May 2020.

of text negativity provides us with a tried and tested financial metric (Aroussi, 2019) for the sentiment that companies—and industries in aggregate—attribute to a certain topic. We further weight the (a) text negativity measure by (b) the number of ‘corona’-keywords per report to reflect how encompassing a company and industry addresses the topic of Covid-19.

Combining both measures into one index has the advantage that the index reflects both how strongly and how frequently firms report about the pandemic. This is why we opt for this simple two-factor metric. The simplicity of the metric makes it more intuitive and easier to interpret than more complicated measures covering multiple variables. In creating the index, we also tried alternative

specifications (e.g., also including the share of reports mentioning Covid-19 per industry), but did not find substantial differences to the simple index consisting of just two variables.

Sentiment analysis. The measure of text negativity is calculated as the share of negative words per sentence that includes at least one ‘corona’-keyword. To identify negative words in SEC reports, we rely on the methodology established by Loughran and McDonald (2011). The researchers have derived a sentiment dictionary of words explicitly applicable to the analysis to SEC filings.

In addition to counting the number of negative words per sentence, we count the total number of words per sentence to calculate the share of negative words per sentence. To derive the overall text negativity measure per industry, we then calculate the average share of negative words per ‘corona’-sentence per day and industry.

Text statistics. The number of ‘corona’-keywords per report measures how often firms mention these keywords. It is a straightforward measure that is meant to assess how relevant the pandemic is for businesses. To calculate the measure, we simply count the number of ‘corona’-keywords per report. While the share of firms that report about ‘corona’ has, in general, increased significantly during the first months of 2020, there are still substantial differences between firms and industries.

From these two measures, we derive the CoRisk-Index as their daily geometric mean for all eight industries and one for the total of all eight industries. When combining the number of ‘corona’-keywords and the text negativity into a composite index, we balance clearness of the new metric and complexity of the data. With regard to this trade-off, our decision to apply a geometric mean is motivated by three aspects. First, for readers of our work and users of our data, who are likely from the field of economics or finance, the geometric mean is an understandable and well-recognised metric (Levy and Robinson, 2006). Secondly, in contrast to other unweighted aggregations, like an arithmetic mean, the geometric mean is less sensitive to outliers, which are very likely to occur in the uneven distribution of keyword counts across documents, as shown in past work (Loughran and McDonald, 2011). Lastly, as our index measures a new phenomenon, by definition, we have little historical knowledge on how to calibrate the two components of our index with weights, which, furthermore, might require constant recalibration over time, as the pandemic evolves. In order to smooth out differences in the daily reporting patterns, we calculate a 14-day moving average (aligned right) of the daily CoRisk-Index values. For the online visualisation, index values are multiplied by 100 to display whole numbers.

Index calculation. Let $K = \{\text{‘corona’}, \text{‘covid’}\}$ be the set of corona keywords, R_t^i the set of reports at time t , which belong to industry i and S_r the set of sentences contained in report r . One sentence s consists of a set of words w , such that $s = \{w_1, \dots, w_{|s|}\}$. We then have $S_t^i = \bigcup \{r \in R_t^i\}$ as the set of sentences contained in reports of industry i at time t . Let $\mathbb{1}_A(x)$ be the indicator function for a set A , given by

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{else.} \end{cases} \quad (1)$$

The number of corona keywords per industry i at point of time t is then given as

$$Key_t^i = \frac{1}{|R_t^i|} \sum_{r \in R_t^i} \sum_{s \in S_r} \sum_{w \in s} \mathbb{1}_K(w) \quad (2)$$

We then determine the set of sentences mentioning at least one of

the keywords as

$$C_t^i = \{s \in S_t^i : k \in s \text{ for some } k \in K\} \quad (3)$$

Then, the text negativity in industry i at time t is given as

$$Sen_t^i = \frac{1}{|C_t^i|} \sum_{s \in C_t^i} \frac{1}{|s|} \sum_{w \in s} \mathbb{1}_{S^-}(w) \quad (4)$$

where $\mathbb{1}_{S^-}(w)$ is the indicator function defined above and S^- are the negative-tone words identified using the Loughran and McDonald (2011) sentiment dictionary. Then, these two measures get combined in a geometric mean, giving the CoRisk-Index for industry i at point of time t :

$$CoRisk_t^i = \sqrt{Key_t^i Sen_t^i}. \quad (5)$$

Additionally, we provide two other measures of industry-specific corona-related risk awareness in the report and on the online dashboard: the share of firms mentioning ‘corona’ at least once in their 10-K reports (the count of a simple indicator variable, taking the value one, if a report contains at least one ‘corona’-keyword and zero otherwise), and the share of topic-specific keywords per industry.

Representativeness

Industries. To reflect the differences in coverage between industries and to avoid excessive extrapolation to those industries that are not well represented, we use the COMPUSTAT database (made available via Wharton Research Data Services²⁰) to obtain employee count data from the reporting firms. Through that database, we could obtain the employee count of 4700 of 6400 reporting firms in 2020. These 4700 firms have a total of 44 million employees in 13 sectors of the US economy. The total economy has 150 million employees in these 13 sectors.²¹ Thus, the reporting firms represent around one-third of all employees in the US economy. However, not all sectors are well represented (see Fig. 7). In eight sectors, the firms filing to the SEC, for which we could obtain data, represent at least 22% of all US employees in that sector. Therefore, we limit the CoRisk-Index to these eight industries.

Seasonality of reporting. The CoRisk-Index reports corona-specific risks on a daily basis. This time series nature makes the index prone to potential biases, if the industries tend to have vastly changing reporting schedules during the calendar year. We investigate the seasonal reporting pattern in Fig. 8. The figure shows the share of reports filed to the SEC per quarter in the period 2014 to 2018. A clear seasonal pattern can be identified in

	Section	SEC Employees (million)	US Employees (million)	Share (%)
1.	Mining	0.7	0.75	93
2.	Manufacturing	12.5	15.7	80
3.	Information	1.7	2.8	62
4.	Wholesale & Retail	10.3	19.7	52
5.	Finance	4.2	10.8	39
6.	Transportation & Utilities	3.4	9	38
7.	Prof. & Business Services	6	19.6	31
8.	Leisure & Hospitality	3.3	14.6	22
9.	Other Services	0.4	7.6	5
10.	Education & Health	1.4	35.9	3.8
11.	Construction	0.3	11.4	2.9
12.	Agriculture	0	2.4	0
13.	Public Administration	0	7.2	0

Fig. 7 Number of employees working in different sectors of the US economy and share of employees working in firms that report to the SEC. In eight out of 13 sectors, the firms that file to SEC represent at least 22% of all employees in that sector.

Fig. 8 Share of reports per quarter of the eight different industries. There is some seasonality, which is mainly driven by Finance's large number of reports in the first quarter. The other industries do not show substantial variation within a calendar year.

the first quarter of each year. However, this seasonality appears to be driven mostly by the Finance industry. This sector reports substantially more in the first quarter. Finance has reported relatively low corona-related risks in the first quarter of 2020, despite the high number of reports. Thus, the CoRisk-Index does not seem to be biased towards those industries that file more reports in a given quarter. The other sectors do not show substantial seasonal patterns. Thus, we conclude that the results presented in the report are not substantially biased due to seasonal reporting patterns.

Granger causality test

Methodology. Relying on the Granger causality test, we evaluate if there is a statistically significant time-lagged correlation between the CoRisk-Index and economic indicators such as the unemployment rate and the S&P 1200 Global Index as well as the VIX index as a recognised measure of market volatility. The null hypothesis for Granger causality is that no such correlation exists. In other words, the hypothesis states that the CoRisk-Index does not 'Granger-cause' any of the economic indicators. Granger causality quantifies the degree to which one time series holds informative value to forecast another time series based on a statistical hypothesis test. Intuitively, Granger causality tests whether the prediction of variable Y based on its own past values and another variable's X history as a regressor is more accurate than predicting Y only autoregressively (based on its own history alone). Our setting fulfils all requirements for testing Granger causality. The 10-K filings have been issued prior to the observed effect and supposedly hold unique informative value. To focus on relative instead of absolute changes, we normalise all time series by taking the first derivative. Our time series can be regarded stationary, free from regular, seasonal trends. We verify stationarity of all time series using the KPSS test with a significance level of $\alpha = 0.05$.

Mathematical statement. Let y and x be stationary time series, e.g., in our case, y is the industry-specific unemployment rate and x

the CoRisk-Index.²² To test the null hypothesis that x , the CoRisk-Index, does not Granger-cause y (unemployment, VIX, S&P 1200), one first finds the proper lagged values of y to include in a univariate autoregression of y :

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_m y_{t-m} + \text{error}_t.$$

Next, the autoregression is augmented by including lagged values of x :

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_m y_{t-m} + b_p x_{t-p} + \dots + b_q x_{t-q} + \text{error}_t.$$

One retains in this regression all lagged values of x that are individually significant according to their t-statistics, provided that collectively they add explanatory power to the regression according to an F -test (whose null hypothesis is no explanatory power jointly added by the x 's). In the notation of the above augmented regression, p is the shortest, and q is the longest, lag length for which the lagged value of x is significant. The null hypothesis that x does not Granger-cause y is accepted if and only if no lagged values of x are retained in the regression.

Limitations. It is important to acknowledge several limitations of the VAR-based Granger causality test. First of all, two variables X and Y may reveal Granger causality, even though they might be confounded by a third, unknown variable. The CoRisk-Index is of course not the cause for any recorded fluctuations in economic indicators, but merely holds predictive signal or 'Granger-causes' the observed economic effects (symptoms). To this end, the CoRisk-Index serves as a sufficiently representative proxy of the actual, but hardly quantifiable causes of the pandemic such as work loss, planning uncertainty and restrictions. Secondly, we operationalise our multivariate Granger causality test by fitting a VAR model. VAR is a parametric model incapable of testing non-linear correlations and correlations in higher order moments.

Fig. 9 Comparison of quarterly GDP change (upper panel), unemployment rates (central panel), and share of negative words in 10-K reports (lower panel). The figure exemplifies that the sentiment of 10-K reports tend to correlate with overall macroeconomic developments.

Historical robustness. The findings provided in the report rely on the assumption that the firm reports filed to the SEC contain information that is reflecting actual economic circumstances, and not just artefacts of specific wording, maybe due to trends, herding or currently relevant ‘hot’ topics. In order to investigate this, we have examined historical SEC filings data (Loughran and McDonald, 2011)²³ and calculated the share of negative words (one of the two key components of the CoRisk-Index) for these filings in two ways.

First, we calculated the share of negative words per report in all 152,694 reports that have been filed between 2000 and 2018. To compare this historical text negativity with macroeconomic data, we aggregated the per-report negativity score for all reports filed in one quarter and compared this to US quarterly unemployment rates and quarter-to-quarter changes in GDP.²⁴ Figure 9 shows the results. During the two recessions (grey bars) that happened in the period from 2000 to 2018, the GDP (upper panel) dropped significantly and the unemployment rate increased (central panel).²⁵ Correlated with these overall macroeconomic developments, the share of negative words in 10-K reports (lower panel) increased in these periods. In periods of economic recovery (lowering unemployment rates in 2004 to 2008 and in 2010 to 2012), the overall text negativity decreased.

This observation provides supportive evidence that textual data extracted from SEC filings contains information that is correlated with the changing real-world economic circumstances. However, the overall text negativity is a highly aggregated measure, not comparable to the fine-grained data on specific topics that are covered by the text analysis of sentences mentioning specific keywords, such as Covid-19. Moreover, during the course of the

past 20 years, the length of 10-K reports has increased substantial, which might influence text mining measures on the report level.

Data availability

All raw data and results are published on an interactive online dashboard. It is being made available on <http://oxford.berlin/corisk>. The dashboard provides an interactive visualisation of the main industry-specific statistics described in the report: (a) the CoRisk-Index, (b) the share of negative words per corona-sentence in the panel ‘Text Sentiment’, (c) the share of reporting firms per industry in the panel ‘Industry View’, and (d) the share of topic-specific keywords per corona-sentence in the panel ‘Topic Heatmap’. Additionally, we have established an easy-to-use filter and download option. It allows researchers, journalists, and the general public to get full access to both the weekly aggregated index data and the raw data on the level of individual reports.

Code availability

In addition to the online dashboard, all code, data and a detailed documentation is available on our Github page (<http://github.com/Braesemann/CoRisk>).

Received: 22 September 2020; Accepted: 22 December 2021;

Published online: 02 February 2022

Notes

- 1 The VIX provides a recognised measure of market volatility on which expectations of further stock market volatility in the near future might be based. Figure 3A shows the correlation of the CoRisk-Index and VIX.
- 2 The German IFO Index, a globally established measure of business climate, requires a monthly survey among 7000 participants with multiple question modules from firms in manufacturing, construction, wholesaling and retailing, who are requested to assess their current business situation as well as their business outlook for the coming six months (Sauer and Wohlrabe, 2018).
- 3 In the paper, we use the term sector and industry interchangeably; however, we want to emphasise that the level of analysis is two-digit sectors of the U.S. economy, according to the North American Industry Classification System. As all raw data are being made publicly available, the analysis could be reproduced on the level of individual firms or four-digit industries and sub-sectors. Details about the data collection are presented in the Methods section ‘Data collection’.
- 4 The data and interactive dashboard can be found at <http://oxford.berlin/corisk/>.
- 5 For details on the VIX: <https://www.investopedia.com/terms/v/vix.asp>.
- 6 Figure 3A shows that the CoRisk-Index anticipates unemployment on both the general economic and industry-specific level; and that it correlates with established risk indicators such as the VIX.
- 7 For readability, the figure shows only two selected industries per panel.
- 8 We have conducted the Granger causality tests on monthly aggregated data as unemployment rates are only available on the monthly level.
- 9 The data and interactive dashboard can be found at <http://oxford.berlin/corisk/>.
- 10 <https://www.sec.gov/Archives/edgar/full-index/>.
- 11 <https://sec-api.io/>.
- 12 <https://www.sec.gov/Archives/edgar/full-index/2020/QTR2/crawler.idx>
- 13 E.g., <https://www.sec.gov/Archives/edgar/data/1084869/0001437749-20-002005-index.htm>.
- 14 Information about the SIC standards used by the SEC can be found here: [http://www.secfinfo.com/\\$/SEC/NAICS.asp](http://www.secfinfo.com/$/SEC/NAICS.asp).
- 15 <https://www.census.gov/eos/www/naics/concordances/concordances.html>.
- 16 <https://github.com/Braesemann/CoRisk>.
- 17 <https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2017>.
- 18 <https://www.sec.gov/fast-answers/answers-form10k.htm>.
- 19 <https://www.sec.gov/news/press-release/2020-53>.
- 20 <https://wrds-web.wharton.upenn.edu/wrds/support/Data/>.
- 21 <https://www.statista.com/statistics/200143/employment-in-selected-us-industries/>.
- 22 The mathematical statement is, to a large extent, an excerpt from the Wikipedia article on ‘Granger causality’: https://en.wikipedia.org/wiki/Granger_causality
- 23 <https://srafi.nd.edu/crsp-flat-file/>.
- 24 The macroeconomic data has been collected from: <https://research.stlouisfed.org/econ/mccracken/fred-databases/>.

25 Recession periods from <https://www.nber.org/cycles.html>.

References

- Aroussi R (2019) yfinance: Yahoo! Finance market data downloader
 Baker SR et al. (2020) The unprecedented stock market impact of COVID-19. Working Paper 26945, National Bureau of Economic Research. <https://doi.org/10.3386/w26945>. Series: Working Paper Series
 Blei DM (2012) Probabilistic topic models. *Commun ACM* 55:77–84. Publisher: ACM New York, NY, USA
 Bochkay K, Joos PR (2021) Macroeconomic uncertainty and quantitative versus qualitative inputs to analyst risk forecasts. *Account Rev* 96:59–90
 Buetre B, Kim Y, Tran QT, Gunasekera D (2006) Avian influenza: potential economic impact of a pandemic on australia. *Australian Commodities: Forecasts and Issues* 13:351. Australian Bureau of Agricultural and Resource Economics
 Cerdeiro DA et al. (2020) World seaborne trade in real time: a proof of concept for building ais-based nowcasts from scratch. Tech Rep, International Monetary Fund
 Chouliaras A (2015) The pessimism factor: SEC EDGAR form 10-K textual analysis and stock returns. SSRN Scholarly Paper ID 2627037, Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.2627037>
 Davis SJ, Hansen S, Seminario-Amez C (2020) Firm-level risk exposures and stock returns in the wake of covid-19. Tech Rep, National Bureau of Economic Research
 Deb P, Furceri D, Ostry JD, Tawk N (2021) The economic effects of covid-19 containment measures. *Open Economies Review*, 1–32. Publisher: Springer
 del Rio-Chanona RM, Mealy P, Pichler A, Lafond F, Farmer D (2020) Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy* 36:Supplement_1, S94–S137, Publisher: APA
 Dyer T, Lang M, Stice-Lawrence L (2017) The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation. *J Account Econ* 64:221–245
 Fama EF (1965) The behavior of stock-market prices. *J Bus* 38:34–105
 Fama EF (1980) Agency problems and the theory of the firm. *J Polit Econ* 88:288–307
 Fama EF (1991) Efficient capital markets: Ii. *J Finance* 46:1575–1617
 Fernandes N (2020) Economic Effects of coronavirus outbreak (COVID-19) on the world economy. SSRN Scholarly Paper ID 3557504, Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.3557504>
 Goodell JW, Huynh TLD (2020) Did Congress trade ahead? Considering the reaction of US industries to COVID-19. *Finance Res Lett* 101578, <https://doi.org/10.1016/j.frl.2020.101578> (2020)
 Hassan TA, Hollander S, Van Lent L, Tahoun A (2019) Firm-level political risk: Measurement and effects. *Q J Econ* 134:2135–2202
 Hassan TA, Hollander S, van Lent L, Tahoun A (2020). Firm-level exposure to epidemic diseases: Covid-19, SARS, and H1N1. NBER Working Paper, National Bureau of Economic Research
 Keogh-Brown MR, Smith RD, Edmunds JW, Beutels P (2010) The macroeconomic impact of pandemic influenza: estimates from models of the United Kingdom, France, Belgium and The Netherlands. *Eur J Health Econ* 11:543–554
 Kinne J, Krüger M, Lenz D, Licht G, Winker P (2020) Corona-Pandemie betrifft Unternehmen unterschiedlich. Working Paper, ZEW-Leibniz-Zentrum für europäische Wirtschaftsforschung Mannheim GmbH
 Kogan S, Levin D, Routledge BR, Sagi JS, Smith NA (2009) Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies-NAACL '09*, 272, <https://doi.org/10.3115/1620754.1620794>. Association for Computational Linguistics, Boulder, Colorado
 Lazer D et al. (2009) Computational social science. *Science* 323, 721–723, <https://doi.org/10.1126/science.1167742>
 Levy H, Robinson M (2006) Stochastic dominance: investment decision making under uncertainty, vol. 34. Springer
 Lewis D, Mertens K, Stock JH (2020) U.S. economic activity during the early weeks of the SARS-Cov-2 outbreak. NBER Working Paper 26954, National Bureau of Economic Research. <https://doi.org/10.3386/w26954>. Series: Working Paper Series.
 Lopez-Lira A (2021) Why do managers disclose risks accurately? textual analysis, disclosures, and risk exposures. *Econ Lett* 204:109896
 Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J Finance* 66:35–65
 Ludvigson SC, Ma S, Ng S (2020) Covid19 and the macroeconomic effects of costly disasters. NBER Working Paper, National Bureau of Economic Research
 Mandl U et al. (2008) The effectiveness and efficiency of public spending. Tech Rep, Directorate General Economic and Financial Affairs (DG ECFIN), European...
 Ozili PK, Arun T (2020) Spillover of COVID-19: impact on the global economy. SSRN Scholarly Paper ID 3562570, Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.3562570>
 Ramelli S, Wagner AF (2020) Feverish stock price reactions to COVID-19. SSRN Scholarly Paper ID 3550274, Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.3550274>
 Richman LD et al. (2019) SEC adopts rules to modernize and simplify disclosure. *J Invest Complian* 20:1–8
 Sauer S, Wohlrabe K (2018) The new ifo business climate index for germany. In *CESifo Forum*, vol. 19, 59–64 (München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der...), 2018
 Theile K, Himme A, Hoberg K (2020) Firm-level risk disclosures: effects on the market value of firms during the risk materialization in the case of the covid-19 crisis. Available at SSRN 3680064, Publisher: SSRN
 Wang VX, Xing BB (2020) Battling uncertainty: corporate disclosures of covid-19 in earnings conference calls and annual reports. Available at SSRN 3586085, Publisher: SSRN
 Werner E, DeBonis M, Kane P, Wagner J (2020) House leaders look to expedite \$2.2 trillion relief package but face possibility that one GOP lawmaker may delay passage-The Washington Post. Washington Post
 Wu JT et al. (2020) Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* 1–5, <https://doi.org/10.1038/s41591-020-0822-7> (2020)
 Zacharakis Z (2020) Wirtschaftsmaßnahmen: Was das alles kostet. Die Zeit
 Zhang D, Hu M, Ji Q (2020) Financial markets under the global pandemic of COVID-19. *Finance Res Lett* 101528, <https://doi.org/10.1016/j.frl.2020.101528>

Acknowledgements

We wish to thank Andrew Stephen, Felipe Thomaz, Felix Reed-Tsochas, Vili Lehdonvirta, Scott Hale, Tobias Reisch, Slava Jankin, Otto Kässi, R.-Maria del Rio-Chanona, and Doyné Farmer for valuable feedback on earlier versions of this manuscript. Moreover, we express our gratitude to Andrew Van Dam for the provision of unemployment initial claims data. We are grateful to have had the opportunity to present our work at DIW Berlin, Hertie School of Governance Berlin, UNDP Eurasia, Oxford Department of International Development, the Etla Institute of the Finnish Economy, and the Institute for New Economic Thinking Oxford.

Author contributions

FS and FB designed research and collected data. FS, FB, OT, and NS analysed data. FS, LN, NS, PD, OT, and FB performed research and wrote the paper.

Competing interests

The authors declare no competing interests.

Ethical approval

Not applicable as this study did not involve human participants.

Informed consent

Not applicable as this study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Fabian Braesemann.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022