



City Research Online

City St George's, University of London

Citation: Fang, L. (2023). Inter-industry convergence within financial services and its systemic implications. (Unpublished Doctoral thesis, City, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://city-ref-test.eprints-hosting.org/id/eprint/30241/>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

Inter-industry convergence within financial services and its systemic implications

Application of text-based methods

Lei Fang

Supervisors: Prof. Gianvito Lanzolla
Prof. Andreas Tsanakas
Dr. Rui Zhu

Bayes Business School
City, University of London

This dissertation is submitted for the degree of
Doctor of Philosophy

February 2023

I would like to dedicate this thesis to my dear supervisors and families, especially my little son.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

I grant powers of discretion to the University Librarian to allow the thesis to be copied in whole or in part without further reference to the author.

Lei Fang
February 2023

Acknowledgements

My four-years PhD life in Bayes is like a magical journal. All people, including me, could not imagine the joy and difficulty I came across until witness it. During the four years, I experienced so many important events, just name a few, my little baby boy was born, I found a dream job in Queen Mary University of London, we got Covid, in addition to the completion of this second PhD degree.

Trying a second PhD is never an easy task. That means I left most of my families and friends and came to a new city, where few people knew my past achievements. I kept doubting whether I made the correct decision. Fortunately, I have excellent supervisors, families, and friends. They make me believe that what I have accumulated, e.g., the five-years PhD study in China, will never be wasted, and it will play an important role someday.

First and foremost I am extremely grateful to my supervisors, Prof. Andreas Tsanakas and Prof. Gianvito Lanzolla, for their invaluable advice, continuous support, and patience during my PhD study. Their intelligence and respectable personality have a persistent impact on me in all the time of my academic research and daily life. I sincerely thank them for spending numerous hours on this thesis. I will never forget the warm encouragement they gave me when I experienced failure on the job market. Without their help, it is impossible for me to find my current job.

It is challenging to complete PhD study with a small baby. Luckily, I have my families, who support all my decisions, keep encouraging me, and help me without hesitate when I need support. I would like to thank my little boy, especially, for being such an angel baby rather than a trouble maker. He resolves all my negative emotion and is my biggest source of happiness.

This endeavour would not have been possible without Dr Rui Zhu. Rui is not only my supervisor, but also a perfect friend. She tried her best to help me through the difficulties as well, and I gained a tremendous relief from her during the four years. Chapter 2 is completed with her big contribution, and I learnt and enjoyed a lot during the numerous meetings.

Abstract

Over the last several years, a trend of convergence within financial services has been observed. This trend, often linked to regulatory and business dynamics, has manifested, for instance, in insurers offering banking products and venturing in asset management. The antecedents and consequences of convergence are still only partially understood. The goals of this research are threefold. First, I expand on the understanding of the antecedents of convergence in financial services by considering factors that thus far have been overlooked. Second, I develop specific and actionable measures of convergence based on text analysis. Third, I study the systemic implications of convergence through the lens of systemic risk.

In the first study, I build a framework to measure text similarity at different aggregation levels, and propose novel measures. I reveal empirically the link between convergence in management attention and systemic risk in the banking industry in the context of the Covid-19 Pandemic: convergence of banks' attention to insurance industry increases across-industry systemic risk, while focusing on industry-idiosyncratic issues reduces within-industry systemic risk. Although a note of caution should be made about the generalizability of such results, my findings demonstrate that text-based methods can offer insights in predicting systemic risk.

In the second study, I borrow from management literature and the Attention-Based View to conceptualize convergence at the level of management attention. I study the impact of digitalization on the context within which management attention is situated. I develop hypotheses on the link between digitalization, management attention, and the specific focus of management attention. I test my hypotheses through advanced text-based methods in the context of the 10-K reports published by insurers, banks and pharmaceutical companies from 2006 to 2018. Analyses largely confirm that digitalization is positively related to convergence in management attention.

In the third study, I complement the second study by focusing on the macro business environment in which management attention is situated – i.e., macro business trends; threat of substitution; and management fashions. My goal is to disentangle, theoretically and empirically, the relative role of the three antecedents in the convergence of management

attention. Through the application of advanced analytical tools, I find that management fashions play a prevailing role in driving convergence in management attention.

Overall, my study contributes to management literature by revealing specific antecedents of convergence in management attention and by providing concrete ways of operationalizing such constructs. It further contributes to the risk analysis literature by showing that factoring in management attention as an antecedent of systemic risk is a promising area of research. Jointly, my findings provide perspectives for policy makers and risk professionals to understand and mitigate exposure to systemic risk.

Table of contents

List of figures	xv
List of tables	xvii
Nomenclature	xix
1 Introduction	1
1.1 Problem statement	1
1.2 Convergence within the financial industry	4
1.2.1 The difference between insurance and banking industries	5
1.2.2 Evidence of convergence across industries	6
1.2.3 Measures of systemic risk in the financial industry	9
1.3 Structure of the thesis	11
2 Text similarity in management research: literature review and proposed convergence measures	13
2.1 Introduction	13
2.2 A literature review: text similarity in management research	14
2.2.1 Finding a benchmark	15
2.2.2 Text vectorization	18
2.2.3 Similarity of document vectors	26
2.3 Text similarity measures at different levels	32
2.3.1 Micro-macro level convergence	33
2.3.2 Macro-macro level convergence	35
2.4 Application of text similarity measures for quantifying industry convergence	37
2.4.1 Micro-macro level convergence	38
2.4.2 Macro-macro level convergence	46
2.5 Choice of text vectorization methods	47
2.6 Discussion and summary	50

3	Digitalization, changes in management attention and convergence across industries	53
3.1	Introduction	53
3.2	Literature review and hypotheses development	56
3.2.1	Contextual antecedents of management attention	56
3.2.2	Hypothesis development: the impact of digitalization on management attention	57
3.3	Data and methods	61
3.3.1	Data sources	61
3.3.2	Dependent variable	63
3.3.3	Independent variables	64
3.3.4	Control variables	69
3.3.5	Regression models	70
3.4	Results	72
3.5	Discussion and conclusions	73
4	Contextual antecedents of industry convergence: text analysis of risk disclosure	77
4.1	Introduction	77
4.1.1	Problem statement	77
4.1.2	A Road map of this chapter	79
4.2	Attention-Based View and antecedents of convergence	80
4.2.1	Antecedents of convergence in risk managers' attention	81
4.2.2	Implications for risk management	82
4.3	Measuring inter-industry convergence	84
4.3.1	Data	84
4.3.2	Construction of the convergence measure	86
4.3.3	Results	87
4.4	Construction of antecedents of convergence	88
4.4.1	Topic modelling of 10-K Business sections	88
4.4.2	Antecedent 1: common trends exposure	89
4.4.3	Antecedent 2: threat of substitution	91
4.4.4	Antecedent 3: management fashions	91
4.4.5	Control variables	92
4.5	Explaining convergence of risk disclosures in terms of contextual antecedents	93
4.5.1	Model selection	93
4.5.2	Variable importance and sensitivity	94
4.5.3	Implication of empirical results	98

4.6 Concluding remarks	99
5 Conclusion	101
References	105
Appendix A Constructing Blurring	125
Appendix B Full lists of keywords of digitalization and business models	127
Appendix C Literature review of text similarity in management research	129
Appendix D Definition of text similarity measures	139
D.1 Overlapping terms	139
D.2 Vector distance / angle	140
D.3 Information-based measures	141

List of figures

2.1	Number of articles adopting text similarity in each year.	15
2.2	Key steps in calculating text similarity.	15
2.3	Percentage of papers using each text vectorization methods	24
2.4	Mapping between management constructs and text vectorization methods. .	25
2.5	Mapping between management constructs and common text similarity measures.	31
2.6	Yearly trend of the average number of similarity methods used in each paper.	32
2.7	The yearly means and 95% confidence intervals of the micro-macro across-industry convergence metrics.	39
2.8	The yearly means and 95% confidence intervals of the micro-macro within-industry convergence metrics.	39
2.9	The relationship of the across-industry convergence against the within-industry convergence at the micro-macro level.	40
2.10	The correlation matrix of the micro-macro across-industry convergence of the four vectorizations.	41
2.11	The correlation matrix of the micro-macro within-industry convergence of the four vectorizations.	42
2.12	Relationship between systemic risk ($\tau_{oppo20_{ki}}$) during the Covid-19 Pandemic and document convergence ($ns_{v_{ki}}$) in 2018, at the across-industry level.	44
2.13	Relationship between systemic risk ($\tau_{own20_{ki}}$) during the Covid-19 Pandemic against document convergence ($na_{sv_{ki}}$) in 2018, at the within-industry level.	45
2.14	Macro-macro across-industry convergence using the first principle component of the original text vectors.	47
2.15	Macro-macro within-industry convergence using the negative determinant of the covariance of the text vectors.	48

2.16	Evaluation of the four vectorizations using accuracy rate (upper left), F1 score (upper right), Kappa score (lower left), and average of 1 - <i>Blurring</i> (lower right).	50
3.1	The impact of digitalization on management attention.	61
3.2	<i>Blurring</i> across insurers, banks and pharmaceutical companies.	65
3.3	Average level of attention to digitalization and business model in each year across companies.	68
4.1	Means and 95% confidence intervals of convergence of risk disclosure in each year by industry.	88
4.2	LDA-generated word clouds, representing the distribution of words within each topic.	90
4.3	ALE plots for the Random Forest regression model.	96
4.4	Importance of each antecedents.	97
4.5	Results of sensitivity analysis based on Shapley values.	98

List of tables

1.1	Examples of convergence between insurance products and corresponding similar products from other financial services industries.	8
2.1	How management scholars derive benchmarks	17
2.2	Examples of Text Vectorization.	21
2.3	Examples of Measures of Document Vectors Similarity.	29
2.4	Definition of similarity at different aggregation levels.	33
2.5	Convergence measures based on text similarity at different levels.	37
3.1	Industries and number of observations (firm/year 10-K reports)	62
3.2	Accuracy rates of classification	64
3.3	List of seed keywords on digitalization and business model	67
3.4	Descriptive statistics (before standardization)	70
3.5	Correlation matrix	70
3.6	Regression results	72
4.1	Antecedents of convergence, implications and risk management response .	84
4.2	Total number of firms for each industry and year.	85
4.3	Average percentage distribution of topics, for documents within each industry.	90
4.4	Results of the performance of different regression models with different parameters	94
C.1	Examples of Management Constructs Directly Measured with Text Similarity.	130
C.2	Temporal and cross-sectional concepts measured by text similarity.	135

Nomenclature

Acronyms / Abbreviations

ABV Attention-Based View

AI Artificial intelligence

ALE Accumulated Local Effects

BIS Bank for International Settlements

CDS Credit Default Swap

CEIOPS European Insurance and Occupational Pensions Authority

CEO Chief Executive Officer

CoVaR Conditional Value at Risk

D2V Doc2vec

ECB European Central Bank

ERM Enterprise Risk Management

EU European Union

FD Ratio Fisher Discriminant Ratio

FDA Fisher Discriminant Analysis

FE Fixed effect panel data regression

FSB Financial Stability Board

IAIS	International Association of Insurance Supervisors
IASB	International Accounting Standards Board
IFRS	International Financial Reporting Standards
IMF	International Monetary Fund
IoT	Internet of Things
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
MACQ	Marginal Attribution by Conditioning on Quantiles
MSE	Mean Squared Errors
NAIC	National Association of Insurance Commissioners
OLS	Ordinary Least Squares regression
Param	Parameter
PC	Property and casualty
Pharma	Pharmaceuticals
RE	Random effect panel data regression
SEC	U.S. Securities and Exchange Commission
SVD	Singular Value Decomposition
TFIDF	Term frequency-inverse document frequency
TF	Term frequency
VaR	Value at Risk
WMD	Word Mover's Distance
XOR	Exclusive OR

Chapter 1

Introduction

1.1 Problem statement

Convergence in management attention across industries is a common phenomenon (Kim et al., 2015; Kwon et al., 2020), particularly visible in financial services (ECB, 2020). The trend of convergence was most evident before the breakout of the Subprime Mortgage Crisis in 2008 (Cummins and Weiss, 2009; Weill, 2009). For example, according to Cummins and Weiss (2009), “one of the most significant economic developments of the past quarter century has been the convergence of the previously separate segments of the financial services industry”. Although the trend was lessened by the Crisis, such convergence is still happening both in the banking industry (e.g., Tziogkidis et al., 2020) and insurance industry (e.g., Cummins and Rubio-Misas, 2022; Turk, 2015).

In the academic literature, the topic of industry convergence has been studied in several disconnected literature streams, e.g., insurance (Cummins and Weiss, 2009), finance (Casu and Girardone, 2010; Degl’Innocenti et al., 2017), risk management (Bushman et al., 2017), law (Turk, 2015), etc., and our understanding of the antecedents and consequences of inter-industry convergence is still partial. In management literature, the topic of industry convergence per se has not been a prevailing one. Yet, management literature provides theoretical lenses to explore antecedents of convergence in a broader way than the one performed by finance and risk literature. Specifically, I maintain that management literature, especially the Attention-Based View (Ocasio, 1997), can help to nail down a more nuanced understanding of the antecedents of convergence.

Convergence in management attention across firms and industries matters because similarity impacts on allocation of organizational resources – e.g., management time, financial resources, technology, talent. At an aggregate, more similarity might imply more allocation of similar resources to similar task hence potentially creating systemic risk, even though the

initial purpose of financial integration is to promote competition and efficiency. On the other hand, dissimilarity might imply diverse allocation of resources hence lower systemic risk. As such, understanding the evolution of the level of similarity between firms and its drivers is important as a proxy to understand the decrease / emergence of systemic risk.

One of the most widely used definitions of **systemic risk** is the one jointly proposed by IMF, FSB and BIS in the report to G20 (IMF et al., 2009): *The risk of disruption to the flow of financial services that is (i) caused by an impairment of all or parts of the financial system; and (ii) has the potential to have serious negative consequences for the real economy.* According to this definition, it is important to study systemic risk because it brings a detrimental aftermath, and the risk spreads from the focal financial institutions to formerly healthy ones, or even beyond the financial industry to many other industries and countries.

Systemic risk is not limited to individual firms or a single industry, but has the potential to spread to the real economy. Therefore, systemic risk is related to the concept of **inter-connectedness**, defined as linkages with other components of the system (IMF et al., 2009). Strong interconnectedness can result in firms sharing similar vulnerabilities that expose them to co-movement of extreme downside outcomes (Billio et al., 2012; Bushman et al., 2017). Measuring text similarity within a single industry is not enough to reflect the potential interconnectedness, through which systemic risk may spread. In addition to within-industry convergence in management attention, I study the convergence *across industries*. My study is not limited to individual firm level, and also includes the industry level.

To predict systemic risk, I use **text data** in financial disclosure to study the convergence within and across industries, because text data reflect the attention or cognition of managers (Dutt and Joseph, 2019) and thus antecedents of risk, while other numerical measures based on financial data are mostly ex post. Recent literature demonstrates that qualitative discussions in financial reports form a rich source of information that can be extracted using natural language processing techniques (e.g., Bao and Datta, 2014; Bushman et al., 2017; Hoberg and Phillips, 2016; Loughran and McDonald, 2016).

I choose insurance and banking industry as my research context. After the Subprime Mortgage Crisis in 2008, increasing attention has been paid to industry convergence in the financial industry (e.g., Bhattacharya et al., 2020; Cummins and Rubio-Misas, 2022; Cummins and Weiss, 2009; Turk, 2015). I pay special attention to the insurance industry when studying systemic risk, because of the standard “AIG-as-outlier” view in the law and regulatory literature, despite the tragedy of AIG in the Subprime Mortgage Crisis in 2008: “The (traditional) insurance industry proved to be extremely resilient during the recent crisis and hence should be more a source of positive examples and solutions for other industries rather than be subject to regulatory rules invented for different institutions and industries

such as banking” (Geneva, 2012). After more than one decade, a new crisis triggered by the Covid-19 Pandemic is happening. In the context of a new crisis, it worth a recheck on the relationship between industry convergence and systemic risk .

The goal of this research is threefold. First, I expand on the understanding of the antecedents of convergence by considering factors that thus far have been overlooked. Second, I develop specific and actionable measures of convergence. Third, I study the systemic implications of convergence through the lenses of systemic risk.

In Chapter 2, I review management literature using text similarity, and discuss the key steps to measure management constructs using text similarity: finding appropriate benchmarks, text vectorization, and measuring similarity. I evaluate different methods at each step, and study the pattern and trend of management constructs and corresponding text similarity methods. In the literature, there are few similarity measures of an individual document against a class of documents, or across classes and within a class. I fill the gap by building a framework of text similarity measures at different aggregation levels, and proposing new measures. Moreover, I discuss how to choose vectorization methods by evaluating the performance of each vectorization method in classification. I apply them to insurance and banking text data as illustration, and find that convergence in management attention can predict systemic risk during the Covid-19 Pandemic: the convergence of a bank’s attention to the insurance industry leads to high across-industry systemic risk, and focusing on banking-industry-idiosyncratic issues reduces its within-industry systemic risk.

In Chapter 3¹, through the lens of the Attention-Based View, I conceptualize convergence at the level of management attention. I show that digitalization systematically changes the contextual antecedents of management attention and demonstrate, theoretically and empirically, that these changes lead to convergence in management attention across firms and industries. Specifically, I reveal that management attention becomes less firm-idiosyncratic, embracing concerns and priorities that span firm and industry boundaries; and becomes focused on business models and on their rewiring in digital ecosystems. I test my hypotheses in the context of banks, insurers and pharmaceutical companies (934 banks; 221 insurers; 223 pharmaceuticals) on text data from the Business and the Risk Factors sections of their 10-K reports (2006 to 2018). 10-K reports are regulatory financial reports submitted to U.S. Securities and Exchange Commission by all the listed companies. I develop a novel text-based measure of convergence in attention, using machine learning and text classification techniques. Model estimates find support for the hypothesis that digitalization is linked

¹This chapter was presented in the 80th annual meeting of the *Academy of Management* in 2020, and the 2nd *Insurance and Data Science Conference* in 2019.

to convergence in management attention in firms across industries, and that this effect is mediated by the attention to business models and the rewiring in digital ecosystems.

Chapter 4² focus on the contextual antecedents of the convergence. I construct a new measure of firm-level inter-industry convergence of risk managers' attention, calculated from text data of insurers' and banks' risk disclosures (10-K report). This convergence measure is based on the silhouette values of vector representations of individual insurers' and banks' risk disclosure documents. Documents are represented as vectors by Doc2vec embeddings. Hence, the measure represents the extent to which a given company's report (e.g. an insurer) is semantically close to those in the industry it does not belong to (e.g. banking). In addition, I aim to disentangle the strategic antecedents of such convergence in risk managers' attention. Specifically, I consider the common trends in the macro environment, and threat of substitution. Text-based measures of those two antecedents are constructed, using topic modelling via Latent Dirichlet Analysis of the Business sections of 10-K reports. Furthermore, I consider risk management fashions, which I measure by the convergence of reports to those in a third industry (in this case, pharmaceutical). The impact of those variables on inter-industry convergence of risk managers' attention is quantified via Random Forest regression. Sensitivity analysis allows us to evaluate the differential impact of the different antecedents of convergence, showing that management fashions have dominant explanatory power. Through this analysis, the notion of systemic risk is expanded and re-conceptualised, to consider the confluence of business and institutional risks.

In the next section of this chapter, I first show that insurance and banking are originally different, but there is phenomenological evidence that convergence in financial services is real and manifests in different ways. Then, I review three typical categories of systemic risk measures from literature.

1.2 Convergence within the financial industry

Insurance and banking industries are traditionally different industries, but evidence shows that they are converging, although this was interrupted by the Subprime Mortgage Crisis. In this section, I review literature on convergence between insurance and banking industry, so as to provide (a) background knowledge of the traditional difference between the two industries, (b) historical evidence of the convergence within the financial industry, and (c) a summary of common systemic risk measures in the financial industry.

²This chapter was presented in *Actuarial Science Seminar Series* held by London School of Economics in 2021.

1.2.1 The difference between insurance and banking industries

Insurance is originally different from banking services in the following aspects.

Aims and functions of products. From the view of consumers, the aims and functions of insurance products are different from banks, because: (a) they protect the investor from *well-identified* and *exogenous* risk sources, leading to sudden and considerable wealth falls; (b) they cover *losses of large scale*, which are larger than the wealth accumulated with the help of other financial services; (c) typical insurance products hedge *discontinuous* risks, e.g., death, unexpected survival to a specific time, sickness, disability, loss of a key property, or financial loss from an accident, while most other financial services aim at hedging continuous risks, e.g., market risk (Beltratti and Corvino, 2008).

Instead, banks provide financial intermediation by matching the assets of depositors with investors. Depositors aim at getting a return from banking products, and investors aim at the capital they needed for the investment.

Operational structures. Compared to banks, insurers (a) pay more attention to *demographic factors*, and operate based on the law of large number (Avraham, 2012); (b) *lack secondary market* where insurance contracts can be traded, though insurers can sell their risk through securitization (Cowley and Cummins, 2005); (c) *do not rely so strongly on external capital markets*. Instead, they manage risk through hedging in their own balance sheets, e.g., mortality and longevity risk hedge each other.

Funding structures. Insurers' balance sheets mainly include *short-term liquid assets* from premiums that are paid with relatively predictable pattern (paid upfront or termly), while their *liabilities are long-term and illiquid*, mainly future claims owed to policyholders (Acharya and Dawsonera, 2011; IAIS, 2011). Banks, on the contrary, have short-term liquid liabilities, mainly deposits which can be withdrawn on demand, and long-term illiquid assets, mainly loans (Mishkin, 2007). Such difference in mismatching of assets and liabilities leads to different risks, which will be discussed later.

Legal features, regulation and accounting practices. Insurers and banks are subject to different regulations and accounting practices. For example, insurers usually adopt cost-based rules, while banks use market-based (or price-based) rules (Beltratti and Corvino, 2008). This is because insurance contracts have two special legal features (Jackson, 1999; Turk, 2015). First, they cover *contingent liabilities*: the insurer is liable to pay the insured contingent on the actual occurrence of the incidents covered by the policy. Second, the contingent liability is also contingent on "*insurable interest*": insurance contracts are usually limited to compensating the financial lost from incidents that the policyholders or related parties suffer, rather than irrelevant third parties.

Risk. Risks in the banking industry are mainly endogenous. Banking industry risks are usually generated from within the financial market, by the behaviour and confidence of stakeholders within the banking industry (Danielsson and Shin, 2003). For example, the greedy and incautious behaviour of large banks and other financial institutions caused the 2008 Subprime Mortgage Crisis (Kaufman, 2014). On the contrary, risks in the insurance industry are mainly exogenous in the sense that they are triggered by events outside the system, e.g., earthquakes, demographic factors, etc. It is true that the behaviour of stakeholders within the insurance market will impact the vulnerability of the insured to the exogenous shocks, the risks are still more exogenous compared to those in the banking industry.

The biggest risk of banks is credit risk, which is risk of default on a debt that may arise from a borrower failing to make required payments (BIS, 2000). The explosion of credit risk leads to runs of depositors during a recession or an economic crisis of confidence (Acharya and Dawsonera, 2011; Mishkin, 2007). In turn, runs reduce the availability of credit, which further exacerbates the recession or the loss of confidence (Reinhart and Rogoff, 2009). Other major risks faced by banks include market risk, the risk of losses in positions arising from movements in market variables like prices and volatility (BIS, 2003), e.g., equity risk, interest rate risk, currency risk, etc.; and liquidity risk, the risk that for a certain period of time a given financial asset, security or commodity cannot be traded quickly enough in the market without impacting the market price (Amihud et al., 2012). The biggest risk for insurers is from demographic factors (Gründl et al., 2006) or natural disasters. Insurers' traditional funding structure mentioned above is more stable by hedging within the balance sheets, and are less susceptible to runs. In addition, traditional insurance covers independent and idiosyncratic risks which follow stable distributions according to the law of large numbers, and are easier to predict.

1.2.2 Evidence of convergence across industries

Despite the substantial differences mentioned above, there is evidence of convergence between the insurance industry and other financial services industries. According to Dungey et al. (2013), "insurance groups are acting, and will increasingly act, as shadow banks, since they replace traditional banks as lenders". Such convergence is closely related to systemic risk (e.g., Cummins and Weiss, 2009), and continues even after the Subprime Mortgage Crisis (Cummins and Rubio-Misas, 2022; Genetay and Molyneux, 2016; NAIC, 2013c; Thimann, 2015). Evidence of convergence can be found in the following aspects.

Products

Evidence of industry convergence at the product level was evident before 2008, and interrupted by the Subprime Mortgage Crisis.

Insurance products are being re-designed to include financial features, which are typical in products from other financial institutions like banks (Beltratti and Corvino, 2008). Table 1.1 list several typical examples of such insurance products and corresponding similar products from other financial services industries.

For example, *unit-linked insurance products* and *index-linked insurance products* embed financial elements into typical insurance products, and even spawn internally managed mutual funds in European markets (Beltratti and Corvino, 2008).

Guaranteed Investment Contracts (GICs) enable the insured to place funds in the insurer in return for guaranteed interest return. Some forms of GICs allow the principal to be called with short notice, leading to the run-prone characteristics (Turk, 2015; Wallison, 2008).

A special example for life insurers is *fixed-indexed annuities*, providing policyholders with guaranteed minimum payments according to a reference market index. Such products are popular since their first appearance in 1995 (NAIC, 2013a). The guarantee of minimum payments and reliance on external indexes expose life insurers to market risk in the stock market, and proved detrimental to insurers during the Subprime Mortgage Crisis in 2008 (Acharya and Dawsonera, 2011). Another example for life insurers is *securities lending programs*, i.e. lending the life insurers' credit (assets) to over-the-counter investors in return for a fee. Such lending activities are similar to bank loans, and raised large losses in the Subprime Mortgage Crisis (NAIC, 2011).

Property and casualty (PC) insurers sell *surety bonds*, linked with certain construction projects. They cover the risk of failure to meet contractual commitment, which share similar functions with commercial loans offered by banks, e.g., stand-by letters of credit (Wallison, 2008). Surety bonds expose PC insurers to the risk from the real estate and construction industries (Aon, 2013). Another example is *insurance-linked securities* (ILS). ILS operate as follows: an insurer creates a special purpose vehicle to issue securities, which are backed by premium income from a pool of insurance policies (Albertini, 2009). ILS enable insurers to directly access capital markets without turning to reinsurance, exposing insurers to the market risk (Cummins and Weiss, 2009). A further example of PC insurers is the notorious credit default swaps (financial guarantee insurance against the default on low-risk municipal bonds), which are believed to be one of the important triggers of the Subprime Mortgage Crisis (Neale and Drake, 2009; Schich, 2008).

Insurance products	Financial products	References
Unit-linked insurance products, index-linked insurance products	Mutual fund products	Beltratti and Corvino (2008)
Guaranteed Investment Contracts	Bank deposit	Turk (2015); Wallison (2008)
Fixed-indexed annuities, with minimum payment guarantee	Mutual fund products	Acharya and Dawsonera (2011); NAIC (2013a)
Securities lending programs	Bank loan	NAIC (2011)
Surety bond, linked to construction projects	Commercial loans offered by banks, e.g., stand-by letters of credit	Aon (2013); Wallison (2008)
Insurance-linked securities, e.g., Cat bonds	Securities	Albertini (2009); Cummins and Weiss (2009)
Credit default swaps	Derivative sold by investment banks	Neale and Drake (2009); Schich (2008)

Table 1.1 Examples of convergence between insurance products and corresponding similar products from other financial services industries.

Asset management

The second aspect of the convergence is insurers' increasing role as asset managers, which is traditionally one of the main business of banks. The trend of convergence in this aspect has been continuing even after the Subprime Mortgage Crisis, and is still evident recently.

At the end of 2019, U.S. insurers own \$5.6 trillion of credit, which is 10.37% out of the aggregate U.S. outstanding credit (SEC, 2020). The main source of insurers' assets are the premiums paid by policyholders, which are traditionally invested in assets of low risk. Nowadays, insurers also invest in similar ways as other financial institutions (NAIC, 2013b). For example, during the Subprime Mortgage Crisis, the whole insurance industry invested more than \$1 trillion in the housing sector and related derivatives, which was about a fourth of that of the entire banking industry (Foote et al., 2012). That means the traditionally independent insurers were connected to the fragile financial institutions during the Crisis (Eling and Schmeiser, 2010). Even after the lesson of the Subprime Mortgage Crisis, such convergence in investment only slightly weakened (NAIC, 2013c).

Outsourcing is common in insurance asset management, because asset management is complex and insurers pursue higher returns. In Europe, insurance assets outsourced to external managers were around €1.3 trillion (US\$1.5 trillion) at the end of 2019, approximately 16% of industry total general account assets (Cerulli, 2021). Such trend further increases the connection or convergence between insurers and other financial institutions.

Marketing / distribution channels

The “bancassurance” channel is an obvious example of convergence between banks and insurance companies. It is derived from the purposes of one-stop shopping to get wider customer access and convenience (Choudhury and Singh, 2021) and economy of scale. Before the Subprime Mortgage Crisis, “one of the most prominent transformations undergone by the financial services industry has been the emergence and expansion of bancassurance” (Artikis et al., 2008; Genetay and Molyneux, 2016). After the Crisis, the evidence of convergence in marketing and distribution channels is mixed, depending on the location and products (Rubio-Misas, 2022; Sigma, 2007).

Regulation and accounting practices

The convergence in regulation were triggered, interrupted or lessened after the Subprime Mortgage Crisis to deal with the regulatory gap. For example, the *Financial Conglomerate Directive* adopted by G10 with the help of BIS, which is related to regulating financial conglomerates, was triggered by the Crisis (Beltratti and Corvino, 2008); The *International Accounting Standards Board (IASB) insurance project* aiming at homogeneity in the balance sheet evaluation methods of insurers and other financial institutions (IASB, 2007a,b), was interrupted by the Crisis; The *Solvency II* project by European Insurance and Occupational Pensions Authority (CEIOPS), changing basic principles of the regulatory capital requirements of insurers to achieve consistency between insurers and banks, was also interrupted to take the potential systemic risk into consideration (CEIOPS, 2007).

Before the Subprime Mortgage Crisis, in 2005, European insurers, traditionally using cost-based accounting rules, started using *International Financial Reporting Standards* (IFRS 4) involving market-based rules (used by banks) (IFRS, 2005). IFRS 17 will replace IFRS 4 after 1 January 2023. It “separates specified embedded derivatives, distinct investment components and distinct performance obligations from the insurance contracts” (IFRS, 2022), and might drive divergence in regulation.

1.2.3 Measures of systemic risk in the financial industry

In the last subsection, I find evidence of convergence between insurance and banking industry. To study its trend and relationship with other antecedents, the first step is measuring the convergence and the consequent systemic risk.

There are three common categories of systemic risk measures in the financial industry.

The first category of systemic risk measures reflects interconnectedness, which is the prerequisite of contagion of risk. For instance, Billio et al. (2012) measure interconnectedness

based on Principal Components Analysis and Granger-causality networks applied to the monthly returns of hedge funds, banks, brokers, dealers, and insurance companies. Cai et al. (2018) measure bank interconnectedness using the similarity of loan portfolios in different industries or regions. Bushman et al. (2017) study the relationship between tail-risk co-movement and text similarity, which is equal to the cosine similarity between each pair of banks' Business and MD&A sections in 10-K reports. They focus on the text similarity within the banking industry.

The second category of systemic risk measures use conditional stress. They may reflect an individual firm's stress conditional on systemic stress, e.g. Brownlees and Engle (2017)'s capital shortfall of a firm conditional on a severe market decline, using data that are both stock-price-based and financial-account-based, and also Acharya et al. (2012), Acharya et al. (2017), and De Jonghe (2010). They may also reflect the system's stress conditional on an individual firm's stress. For instance, 'CoVaR' (Tobias and Brunnermeier, 2016) reflects the change in the VaR of the financial system conditional on an institution being under distress relative to its median state. As individual stress is related to capital charge or 'systemic importance', Tarashev et al. (2016) uses a game-theoretic concept, the Shapley value, to allocate capital charge to each institution based on its systemic importance.

The third category regards systemic risk as the distress insurance premium, and is calculated using risk-neutral probability of default derived from CDS spread (Chen et al., 2014; Huang et al., 2009). Different from interconnectedness, which emphasizes on the contagion of risk, the other two categories of measures also take the scale of the impairment into account, though such scale is measured ex post. Besides, the three categories usually use market or financial data to measure systemic risk.

The industry convergence studied in this thesis is most closely related to the first category, interconnectedness. Instead of numerical data such as stock return or financial data, I use text data to reflect systemic risk. Text data reflect the attention or cognition of managers. While numerical data, e.g. loan portfolio and financial data, show the aftermaths of certain human behaviors, cognition reflected by text data guides human behaviour, and thus is the antecedent of risk. In other words, text data may reflect systemic risk more timely, if such text data are informative.

Text analysis and natural language processing are widely used in different areas of risk analysis, including risk culture (Agarwal et al., 2019), psychological response to human mortality risk (Bhatia et al., 2021), food safety hazards (Goldberg et al., 2020), risk perception and the public acceptance of new technology (Clothier et al., 2015), railway accident analysis (Cao et al., 2020), and safety-related disruptions (Blanco et al., 2019). Data sources of these researches include news articles, social media posts, and accident analysis reports. I

contribute to this area by applying text analysis to reflect the financial systemic risk across industries.

1.3 Structure of the thesis

Based on the phenomenological evidence of convergence between insurance and banking industry, in the following chapters, I study the antecedents of the convergence and its system-wide consequences.

In Chapter 2, I carry out a review on text similarity, which will be used to measure convergence and different aspects of systemic risk. I find a gap that there are very few measures of similarity of one against a class, or across and within classes. I propose new measures at different aggregation levels, and apply them to our 10-K data as illustration. I find that text-based convergence measures may be useful for predicting systemic risk. In Chapter 3, I show that digitalization drives the convergence of management attention by pervasively changing the contextual antecedents of the convergence. In Chapter 4, I build a framework on the contextual antecedents of the convergence, which reflects different aspects of systemic risk.

Chapter 2

Text similarity in management research: literature review and proposed convergence measures

2.1 Introduction

In this chapter, I carry out a systematic review of text similarity measures used in management research. The motivation is to measure convergence in management attention, which is the antecedents of systemic risk, and explore alternative measures of different aspects of systemic risk based on text similarity.

Evaluating the similarity between two or more documents is a key activity in organizational research that uses on text data (Kobayashi et al., 2018), e.g., in innovation and novelty (Arts et al., 2021; Higham et al., 2021; Kelly et al., 2021; Toubia and Netzer, 2017; von Hippel and Kaulartz, 2021), customer agility (Zhou et al., 2018), firm relationship (Chen et al., 2021; Hoberg and Phillips, 2016), matching or relevance of content (Liu and Toubia, 2018; Liu et al., 2020; Piezunka and Dahlander, 2015, 2019), etc. These studies usually involve large samples of text data, which would make it difficult to adopt traditional qualitative research methods, because the manual coding would be prohibitively labor-intensive (Kobayashi et al., 2018; Schmiedel et al., 2019). Text similarity can be further used in clustering (Bao and Datta, 2014; Timoshenko and Hauser, 2019) and network analysis (Chen et al., 2021, 2019; Hoberg and Phillips, 2016; Netzer et al., 2012).

I find gaps in the literature that (a) there are few proper measures of the similarity of one document against a class of documents, or the similarity between two classes of documents; (b) few articles distinguish between the similarity measures across classes and within a class.

This chapter makes methodological contributions to addressing these practical questions. I propose new text similarity measures that can be used for measuring convergence at different levels of aggregation.

Finally, as illustration, I apply the new similarity measures to insurance and banking text data. I use the text similarity measures to reflect convergence in management attention across and within industries, and use it to predict systemic risk during the Covid-19 Pandemic.

2.2 A literature review: text similarity in management research

In this section, I carry out a systematic literature review on management research adopting text similarity to get an understanding of the methods used and the management constructs represented.

I search the following keywords: text analysis, text mining, natural language processing, topic modelling, in the following journals since 2010:

1. All the 50 journals in Financial Times Research Rank,
2. and all the 4*, 4 and 3 journals on organization studies and all the 4* and 4 journals on general management, ethics, gender and social responsibility, according to Academic Journal Guide 2021.

Then, I read through the articles and manually identify articles using text similarity. In total, I have 67 articles that meet the condition, which are summarized in Appendix C as the original data for further coding.

Figure 2.1 reports the number of papers using text similarity methods in each year. It shows that the total number of management research adopting text similarity is small, indicating that this is a new and under-exploited area. However, over the past decade, the number has been increasing, especially after 2018. The increasing momentum and the fast development of machine learning implies a promising future of this method and urgent need of a guidance on how to apply text similarity in management research.

In the following sections, I review and discuss the key steps in calculating text similarity (see Figure 2.2), namely, finding a benchmark (Section 2.2.1), text vectorization (Section 2.2.2), and measuring text similarity (Section 2.2.3). I discuss pros and cons of the methods, and try to find pattern or trends of the usage of them.

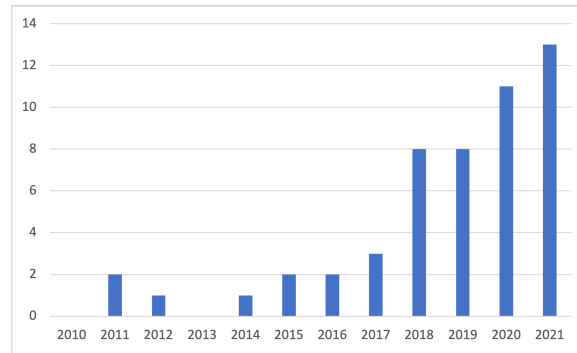


Fig. 2.1 Number of articles adopting text similarity in each year.

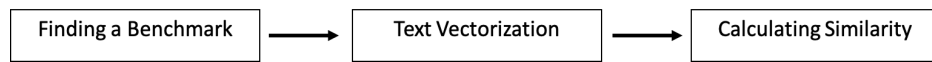


Fig. 2.2 Key steps in calculating text similarity.

2.2.1 Finding a benchmark

First, we need a benchmark, a reference or standard against which the similarity of the focal document (or class of documents) can be calculated. Table 2.1 summarizes different types of benchmarks to measure text similarity. Taking the management construct, innovation, as an example, the benchmark can be all the past patents, idea or content (Arts et al., 2021; Burtch et al., 2021; Kelly et al., 2021). If the focal document is dissimilar enough to the benchmark, the creator is judged to be innovative. The benchmark can also be a list of innovation concepts (von Hippel and Kaulartz, 2021). The creator can be claimed to be innovative if the focal document is similar to the list of concepts. Setting reliable and authentic benchmarks derived in a scientific way is a key step to measure text similarity in such applications and contexts.

In many other articles, text similarity is not used to directly measure the management constructs themselves, but to compare the same management constructs cross-sectionally or temporally (summarized in Table C.2), and the management constructs themselves are measured by methods other than text similarity. In other words, they are measured against other documents in the corpus, rather than against a pre-defined benchmark. To measure such text similarity, the benchmarks are all the other sample documents of different individuals or different periods. I do not discuss the benchmarks of this type in detail considering the simplicity.

As is shown in Table 2.1, the first type of benchmarks is a list of keywords. The list of key concepts can be derived from specialists (Mousavi et al., 2020), the authors (Gallus and Bhatia, 2020), or the authors first deductively get a list of seed keywords, and then expand the list by using word-space models to get semantically similar words (Taeuscher et al., 2021;

von Hippel and Kaulartz, 2021). The last method of getting the list of keywords is more complicated, adopting advanced word-space models, and is more reliable than earlier articles.

Here are some examples of the first type of benchmarks. Gallus and Bhatia (2020) identify a list of gender pronouns, gendered relationship words, and other words describing men and women. They measure the female-typedness of the document by calculating the relative semantic distance between a document and the list. In Mousavi et al. (2020), an answer from an unprofessional is believed to be of high quality if it has a high text similarity with a list of answers derived from specialists. von Hippel and Kaulartz (2021) compare the sample document with a list of innovation concepts, and a high text similarity implies high level of innovation of the document. In the context of crowdfunding, Taeuscher et al. (2021) proxy normative legitimacy by calculating the text similarity between each crowdfunding narrative and a list of keywords on the venture's contribution to the community.

The second type of benchmarks is past or public sources. For example, past / existing content is used to measure innovation, novelty, uniqueness, and consistency specifically. Arts et al. (2021) and Kelly et al. (2021) use past idea reflected by patents as the benchmark to measure innovation, and a low text similarity implies high level of innovation. Similarly, Burtch et al. (2021) compare the users' own past (awarded) content with the focal one, and low text similarity implies high self-novelty. Pan et al. (2019) use existing incumbent firms' product description as the benchmark to measure product uniqueness of new entrants.

Public sources, e.g., officially released public information, can also be a benchmarks. For example, release notes of a sequential innovation of Apple products is used as the benchmark to measure customer agility when compared with customer-generated content (Zhou et al., 2018); Job description is the benchmark to measure experience relevance when compared with applicant's past positions (Sajjadi et al., 2019).

Another source of public content is key articles or books on the topic, which is the benchmark to reflect, e.g., the relevance of academic research to marketing practice (Jedidi et al., 2021).

Sufficient justification is needed when disclosing the chosen benchmarks in addition to building the benchmarks in a scientific way. For example, if lists of keywords are set as the benchmarks, we first deductively get a list of seed keywords, and then inductively expand it to include all semantically similar words using models such as Word2vec (Mikolov et al., 2013a). The initial seed keywords can be created by referring to key articles and books, consulting specialists, discussing with focus groups, or a combination of the three. Good benchmarks are not limited to the benchmarks discussed above. Other benchmarks, such as interviews, search results from Google, etc., may also be appropriate.

Types of benchmarks	Management Constructs	References
<i>Lists of keywords / answers</i>		
Keywords derived from the authors	Female-typedness	Gallus and Bhatia (2020)
Keywords / answers derived from specialists	Quality of idea	Mousavi et al. (2020)
Seed keywords derived from the authors, and then expand the list of keywords by word embedding models	Innovation	von Hippel and Kaulartz (2021)
	Normative legitimacy	Taeuscher et al. (2021)
<i>Past / public sources</i>		
Past content, e.g., collection of all past patents	Innovation	Arts et al. (2021); Kelly et al. (2021)
Users' own past (awarded) content	Self-novelty	Burtch et al. (2021)
Existing incumbents' (e.g., product) description	Product uniqueness	Pan et al. (2019)
Officially released public information, e.g., job description, release notes of a sequential innovation of Apple products	Customer agility	Zhou et al. (2018)
Key articles / books on the topic	Experience relevance Relevance of academic research to marketing practice	Sajjadijani et al. (2019) Jedidi et al. (2021)

Table 2.1 How management scholars derive benchmarks

2.2.2 Text vectorization

As machine learning algorithms operate on a numeric feature space, a requisite step to calculate text similarity is feature extraction, or text vectorization. Text vectorization transforms a document into a vector representation in a defined feature space (Bengfort et al., 2018). All the text vectorization methods are summarized in Table 2.2. I consider text vectorization methods including bag of words, text embedding, score to each category, and topic modelling.

Bag of words

I start with *bag-of-words* vectorization methods, by which the length of a vector is equal to the size of the vocabulary in the corpus. Here the corpus is a language resource consisting of structured set of texts, which are used to do text analysis; and the vocabulary is a list of all the unique terms that appeared in the corpus. The simplest text vectorization method is *term frequency*. With this method, each element of the vector is the count of each word that appears in the document (Manning et al., 2008). We get a sparse vector that uniquely represents each document.

Similarly, for the *one-hot* encoding, instead of the count of each word, each element is the presence of each word in the document, forming a boolean vector for each document.

The two bag-of-words vectorization methods mentioned above treat each word separately, without taking into account its context. *Term frequency-inverse document frequency (TF-IDF)* considers the relative frequency or rareness of terms in the document against their frequency in other documents (Salton and Buckley, 1988). For example, in the financial reports of a bank, terms such as 'loan', 'mortgage', and 'interest' appear more frequently in the reports of banking industry, while terms like 'profit', 'revenue', 'debt' appear frequently throughout the corpus of financial reports across all industries, and are less informative. TF-IDF emphasizes the terms that are very relevant to the banking-specific instance. For each term, the relevance of a term to a document is measured by the scaled frequency of the appearance of the term in the document, normalized by the inverse of the scaled frequency of the term in the entire corpus.

Instead of using the inverse of the scaled frequency of the term in the entire corpus as the weight, other self-invented weighting can also be used to normalize term frequency. Examples of *self-invented weighted term frequency* include Chen et al. (2021), who weight word features in a sentence by their distances to the firm mentioned in a sentence; and Gabel et al. (2019), using a probability distribution density as sampling weighting, and the density is calculated based on the number of times each product occurs in training data.

Another example is *keyword frequency*. The authors pre-define a dictionary of filtered keywords, and each document is represented as the word frequency of each keywords (e.g., Arts et al., 2021; Criscuolo et al., 2017; Theurer et al., 2021; von Hippel and Kaulartz, 2021; Zhou et al., 2018). This method is suitable when we are only interested in the keywords which represent the key features of the text, while other words are irrelevant. Different from other bag-of-words vectorization methods, vectors derived by this method have much fewer features and are less sparse.

Text embeddings

One problem of the above four vectorization methods is that if two documents do not use exactly the same terms but are semantically similar, the similarity between them will be 0. For example, the two sentences 'Andrew Bailey predicts a high inflation rate' and 'The governor of the Bank of England expects soaring prices' will get a low similarity if we use the above four vectorization methods. However, 'Andrew Bailey' and 'The governor of the Bank of England', 'predicts' and 'expects', 'a high inflation rate' and 'soaring prices' are semantically close to each other. To solve the problem, *text embeddings* encode text along a continuous scale with a distributed representation. The resulting document vector is not a simple mapping from term position to term score. The document is represented in an embedded feature space that shows the semantic relationship among terms, rather than a simple mapping from term to term to an element in a sparse vector.

Word2vec is an example of text embedding. The Word2vec algorithm trains word representations based on either a continuous bag-of-words or skip-gram model, such that words are embedded in space along with semantically close words based on their context (Mikolov et al., 2013a,b). Using Word2vec embeddings of words, Word Mover's Distance (WMD) enables us to assess the semantic distance between two documents (Kusner et al., 2015).

We can get the word embeddings by training on the data collected by the users, or by directly using *pre-trained word embeddings*, e.g. GloVe (Pennington et al., 2014).

Based on Word2vec, a document can be represented by taking a simple average of word embeddings corresponding to the words in the document. An alternative to this is *Doc2vec*, an extension of Word2vec. The output is a document vector, representing a whole document rather than a single word. It also attempts to reflect the semantic relationship of terms like Word2vec. Moreover, the paragraph vector takes into account the ordering of words within a narrow context, similar to an n-gram model. The Doc2vec representation is much more effective than a bag-of-words or bag-of-n-grams model because it generalizes better and has

a lower dimensionality, and it performs better than averaging Word2vec (Le and Mikolov, 2014).

By using bag-of-words encoding, the length of each document vector is equal to the size of the vocabulary in the corpus. The dimensionality is usually high, and further dimension reduction is usually performed, which will be discussed later. When applying text embeddings, users choose the dimensionality of the word vectors. There are few guidelines on the optimal number of dimensions that should be used while training word embeddings, and according to a rule of thumb, they are usually between 50 to 300 (Patel and Bhattacharyya, 2017).

Scoring methods

In addition to the above two dimensionality choices, students of text similarity may also decide the length of document vectors by choosing the number of categories (features), and assign a *score to each category*. An extreme case is to assign a *single score* to each document, according to, e.g. the sentiment (e.g., Cookson and Niessner, 2020) or length of the document (e.g., Deng et al., 2021), leading to a dimensionality of one. The *Linguistic Inquiry and Word Count* (LIWC) Dictionary, initially built by Pennebaker et al. (2001), is widely used in sentiment analysis. For each linguistic category, researchers count the relative number of words in the document belonging to each specific category. Other examples of the categories include different dimensions of product uniqueness (Kostovetsky and Warner, 2020), and cultural categories (Song et al., 2018). Compared to other text vectorization methods, the choice of categories involves more human judgement.

Topic modelling

Another common text vectorization method in management research is *topic modelling* (Hannigan et al., 2019), by identifying topics in all the corpus using topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Topic models are a type of statistical models for abstracting topics and thus hidden semantic structures in a corpus. One of the outputs of topic models is the topic distribution vector of each document, which represents a probability distribution over all the topics, i.e. how much the document discusses each topic. Such topic distribution vector can be used to represent each document.

Table 2.2 Examples of Text Vectorization.

Text vectorization methods	Description of the methods	References
<i>Bag of words</i>		
Term frequency	Form a sparse vector according to the count of each word.	Liu et al. (2020), Sajjadi-ani et al. (2019), Glasserman and Mamaysky (2019), Lee and Bradlow (2011), Nam et al. (2017), Netzer et al. (2012), Toubia and Netzer (2017), Abbas et al. (2018), Gorbatai et al. (2021), Kim et al. (2015)
One-hot	Form a sparse vector according to the presence of each word. The value is true (1) if the term exists in the document and false (0) if it does not.	Hoberg and Phillips (2016); Lee et al. (2018)
Term frequency-inverse document frequency (TF-IDF)	Place greater weight on words that are more unique to certain documents and lower weight on words that are common across documents.	Piezunka and Dahlander (2015), Kelly et al. (2021), Pan et al. (2019), Burtch et al. (2021), Abbas et al. (2018)
Self-invented weighted term frequency	Weight word features using self-invented weighting methods.	Chen et al. (2021), Gabel et al. (2019)
Keyword frequency	Pre-define a dictionary of filtered keywords. Each document is represented as the word frequency of each keywords.	Criscuolo et al. (2017), Theurer et al. (2021), Arts et al. (2021), von Hippel and Kaulartz (2021), Zhou et al. (2018)
Continued on next page		

Table 2.2 – continued from previous page

Text vectorization methods	Description of the methods	References
<i>Text embeddings</i>		
Word2vec	The Word2vec algorithm trains word representations, such that words are embedded in space along with semantically close words based on their context	Gallus and Bhatia (2020); Guo et al. (2020); Wang et al. (2021)
Pre-trained word embeddings	E.g, patent embeddings constructed by Google, wherein patent full-texts are embedded in a low-dimensional vector-space	Higham et al. (2021)
Doc2vec	The paragraph vector reflects the semantic relationship of terms, and takes into account the ordering of words within a narrow context	Burtch et al. (2021)
<i>Scoring methods</i>		
Single scores	Assign a score to each document, according to, e.g., sentiment, length etc.	Cookson and Niessner (2020); Deng et al. (2021)
Linguistic Inquiry and Word Count (LIWC) Dictionary	For each linguistic category, count how many words in the text relative to the total number of words in the text belong to each specific category.	Piezunka and Dahlander (2019), Adamopoulos et al. (2018), Akstinaite et al. (2020)
Continued on next page		

Table 2.2 – continued from previous page

Text vectorization methods	Description of the methods	References
Each entry in the vector represents the presence / extent of the content attributes	Code each document into different categories. Represent each document with a vector, with each entry in the vector representing a variable for the presence / extent of the content attributes.	Kostovetsky and Warner (2020), Song et al. (2018)
<i>Topic modelling</i>		
Topic modelling	Identify topics in all the corpus using topic models such as LDA, and subsequently represent each document as a probabilistic representation of these topics.	Taeuscher et al. (2021), Corritore et al. (2020), Adamopoulos et al. (2018), Antons et al. (2019), Mousavi et al. (2020), Jedidi et al. (2021), Bao and Datta (2014), Deng et al. (2021), Burtch et al. (2021), Liu and Toubia (2018), Gątkowski et al. (2020), Buenstorf and Heinisch (2020), Ashtor (2022), Anckaert et al. (2020), Zhou et al. (2018)

Discussion of text vectorization methods

Figure 2.3 shows the percentage of papers using each text vectorization methods. The most widely used vectorization method is the simplest bag of words, with almost half of all the papers adopting it. Among the bag-of-words methods, term frequency is still the most popular, with 17.39% papers adopting it. The least used vectorization method is text embeddings, because it is newer and gains its popularity after 2020.

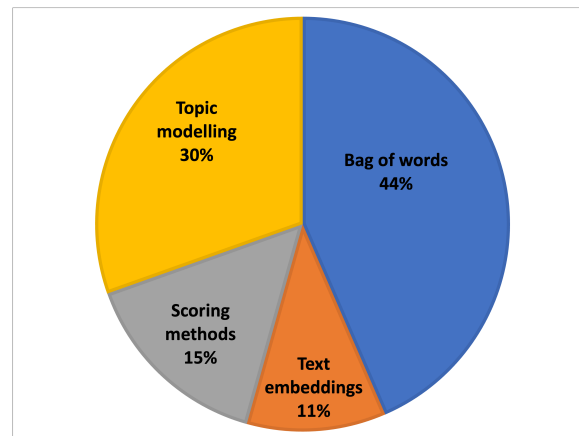


Fig. 2.3 Percentage of papers using each text vectorization methods

There is no obvious pattern in the mapping of management constructs and text vectorization methods, as is shown in Figure 2.4. This implies that all the vectorization methods have pros and cons, and there is no optimally suitable method for each management construct.

Vectors derived from the vectorization methods above, e.g., bag of words, may involve high and sparse dimensions, and therefore dimension reduction is a necessary further step (Zhao and Mao, 2017). A common dimension reduction method is *Singular Value Decomposition* (SVD), applied to the term frequency or TF-IDF vectors, e.g., Wang et al. (2018); Zhou et al. (2018). Archak et al. (2011) use discriminative regression to keep the most important regressors (in their case, the top 20 most popular opinion phrases) to reduce the number of text features.

Although text embeddings vectorization methods have many advantages compared to bag-of-words, e.g., semantic context is considered, the resulting text vector is of reasonable dimension, etc. (Mikolov et al., 2013b), text embeddings have some weaknesses. The biggest weakness is large consumption of computation resources, because they rely on the training of neural networks. The process is slower compared to bag-of-words. Another weakness is the reliance on the third party if pre-trained word embeddings are used, such as the patent embeddings constructed by Google. The training process is usually a black box to management researchers, and the embeddings may be mistakenly applied to inappropriate samples because the users do not fully understand them.

Compared to bag of words and text embeddings, vectorization methods using scoring methods involve much more human interpretation. This type of method is usually used to study emotion, culture, or other concepts that are difficult to quantify.

Topic modelling is a balanced text vectorization method, involving some human interpretation when interpreting each topic, and the dimension of the text vectors is not high. The key

		Bag of words	Text embeddings	Scoring methods	Topic modelling
Management constructs directly measured by text similarity	Innovation, novelty	3	2		1
	Customer agility	1			1
	Quality of idea				1
	Intellectual distance / experience relevance	2			1
	Female-typedness		1		
	Normative legitimacy				1
	Market structure	2			
Cross-sectional or temporal management constructs measured by text similarity	Unusualness	1			1
	Convergence / divergence of opinions / ideas, agreement /disagreement		1	1	1
	Match / relevance of content	2		2	4
	Personality similarity		2	2	1
	Product uniqueness / similarity	3		1	
	Culture congruence / heterogeneity			1	1
	Industry convergence	1			
	Similarity of brands' social tags	1			
	Coherence/changes of identity	1			
	Bibliographic coupling network	1			
	Firm co-benefit relationship	1			
	Relationship between risk types				1
	Distinctiveness between entrepreneurial story and the prototypical				1
	Distinctiveness of employer image or with third-party employer ratings	1			
	The effect of social media advertising content on customer engagement	1			

Fig. 2.4 Mapping between management constructs and text vectorization methods. The number represents the frequency of papers adopting a certain text vectorization method (column titles) corresponding to a certain management construct to be measured in the same paper (row titles). The color becomes darker when the number of papers is larger. Please note that one paper may adopt more than one text similarity measures; and some papers do not involve or clearly disclose the text vectorization methods, so they are not counted in this figure.

parameter when running topic modelling is the number of topics, which are usually set based on the resulting interpretability of the topics or perplexity (Bao and Datta, 2014; Buenstorf and Heinisch, 2020).

2.2.3 Similarity of document vectors

Based on the document vectors derived from the text vectorization methods mentioned above, the next step is calculating text similarity. Common text similarity measures are summarized in Table 2.3, including five types: overlapping terms, vector distance / angle, information-based measures, similarity measures from a third party, and similarity of one document to a group of documents. The formal definitions of all the following text similarity measures can be found in Appendix D.

Overlapping terms

The simplest measure of text similarity measure is counting the *number of words overlapping*, or the count of elements that are both non-zero in the two document vectors. The more overlapping of the number of words, the more similar the two documents are. This method is suitable to documents of similar length.

Jaccard similarity is used to adjust the possible bias derived from various document lengths. It measures similarity between two sample sets, and is calculated by taking the ratio of intersection over union (Jaccard, 1912). When applied to text similarity, it can be, e.g., the ratio of the count of overlapping terms in two documents, divided by the the total size of vocabulary (e.g., Netzer et al., 2012; Powell et al., 2016). We can also replace "terms" with "topics", and the Jaccard similarity become the ratio of the total percentage of overlapping topics in two documents, divided by the the total number of topics (e.g., Adamopoulos et al., 2018). Definition of Jaccard similarity applied in text similarity is shown in Equation (D.4).

Exclusive OR or *XOR* is a logical operation that is true if and only if its inputs *differ* (Simpson, 1974). Applied to text similarity, a value of 1 for the XOR function means that a feature (e.g. the presence of a term) in two document vectors has different values, and 0 the same value. Therefore, a value of 1 implies no overlapping, i.e. dissimilarity, and 0 similarity. A high value of the *mean of the XOR vector* reflects high dissimilarity, and its formal definition is given in Equation (D.7).

Vector distance / angle

For a one-dimension single score representing a document, text similarity can be either the *absolute difference* (e.g., Deng et al., 2021) or *ratio* of the scores (e.g., Akstinaite et al.,

2020). For vectors with more than one dimension, *Euclidean (L2) distance* can be used, as is defined in (D.8).

The most popular document similarity measure is *cosine similarity*, defined as the cosine of the angle between two document vectors, i.e. the dot product of the vectors divided by the product of their lengths, is independent of the length of the documents (Singhal et al., 2001). Its formal definition is in Equation (D.9). This measure is adopted by 27 out of the 67 articles in my reviewed sample publications.

Information-based measures

Measures borrowed from *information theory*, e.g. *entropy*, can also reflect Text Similarity. Entropy is a property that is most commonly associated with a state of disorder, randomness, unusualness, or uncertainty (Daróczy, 1970). Entropy of a random variable is the average level of "unusualness" inherent to the variable's possible outcomes. When applied to text analysis, a document vector is unusual (or distant from other vectors) if it has high entropy. I define entropy applied in text similarity based on Kao et al. (2004)'s definition as is shown in Equation (D.12). A high value of a document vector's entropy implies the document is dissimilar to all the other documents in the corpus.

As a function of entropy, *perplexity* can also be used to measure text similarity. Perplexity is a measurement of how well a probability distribution predicts a sample (Klakov and Peters, 2002). A low perplexity indicates the probability distribution is good at predicting the sample. Applied to text similarity calculation, we replace probability distribution with feature (e.g., term, topic) distribution in a document, and the definition is in Equation (D.14). Consistent to entropy, a high value of perplexity implies dissimilarity or unusualness.

Also based on entropy, *Kullback-Leibler divergence* measures relative entropy, which is a statistical distance between two probability distributions (Kullback and Leibler, 1951). When measuring text similarity, it is the expectation of the logarithmic difference between the feature distribution of two document vectors, as defined in Equation (D.15). As can be seen in Equation (D.15), the measure depends on the choice of the reference vector, and is not symmetric. *Jensen-Shannon divergence* extends Kullback-Leibler divergence to calculate a symmetrical score and distance measure of one distribution from another, as is defined in Equation (D.16). Higher values of the two measures imply larger dissimilarity between two document vectors.

Closely linked to entropy and Kullback-Leibler divergence, *Pointwise Mutual Information* or *Lift* is a measure of association used in information theory, which compares the probability of two events occurring together to what this probability would be if the events were independent (Church and Hanks, 1990). Applying this idea to text analysis, it draws on the

intuition that the best way to weigh the association between two features is to ask how much more the two co-occur in a corpus than we would ex ante expected them to appear by chance (Jurafsky and Martin, 2021). It is worth noting that instead of comparing two document vectors, Pointwise Mutual Information measures the similarity between two features, and the features are usually terms. For example, Netzer et al. (2012) use the measure to calculate similarity between terms (car brands) and any terms that appear frequently in a forum (e.g., safety-related words).

Similarity measure from a third party

Di and Laux (2021)'s similarity measure is directly from a third party, *RavenPack's Relevance Score*. Hafez (2011, p.31) provides the most detailed introduction of RavenPack's Relevance score, which is a relevance score between 0 (low relevance) and 100 (high relevance) indicating how strongly the firm is related to the underlying news story. RavenPack does not disclose the way they calculate the score to the public.

Similarity of one document to a group of documents

All the above similarity measures reflect pairwise document similarity, i.e. how similar one individual document vector is to another individual counterpart, or to a benchmark. Sometimes we also want to measure the similarity of one individual document to a group of documents. For example, we want to study how closely one firm is related to a whole industry. we can use either the average distance between the focal document vector and all the document vectors within the group (Bushman et al., 2017), or the difference between the individual and a prototype document vector that represent the group. For instance, Taeuscher et al. (2021) use the difference between the individual and prototypical topic distribution, which is the total absolute value of the difference between (a) the individual's topic distribution vector and (b) the *average* of those of all the documents belonging to the same the category. Liu et al. (2020) name the prototypical or standard document vector as *Centroid*.

Borrowed from machine learning clustering analysis, *silhouette value* reflects how similar a document vector is to its own cluster (cohesion) compared to other clusters of documents (separation) (Rousseeuw, 1987). Silhouette value is a normalized version of the intra-class distance, considering the inter-class distance relative to the intra-class distance, and its definition is given later in Section 2.3.1. This measure will be adopted in Chapter 4.

Table 2.3 Examples of Measures of Document Vectors Similarity.

Measures of Document Vectors Similarity	References
<i>Overlapping terms</i>	
Number of words overlapping	Kostovetsky and Warner (2020); Liu et al. (2020); Taeuscher et al. (2021)
Jaccard similarity	Adamopoulos et al. (2018); Netzer et al. (2012); Powell et al. (2016)
The mean of Exclusive OR (XOR) vectors	Lee et al. (2018)
<i>Vector distance / angle</i>	
Absolute difference or ratio of Scores	Akstinaite et al. (2020); Deng et al. (2021)
Euclidean (L2) distance	Sajjadiani et al. (2019); Timoshenko and Hauser (2019)
Cosine similarity	Abbas et al. (2018); Arts et al. (2021); Ash-tor (2022); Buenstorf and Heinisch (2020); Burtch et al. (2021); Chen et al. (2019); Criscuolo et al. (2017); Deng et al. (2021); Gallus and Bhatia (2020); Gątkowski et al. (2020); Gorbatai et al. (2021); Guo et al. (2020); Higham et al. (2021); Hoberg and Phillips (2016); Kobayashi et al. (2018); Lee and Bradlow (2011); Liu and Toubia (2018); Liu et al. (2020); Mousavi et al. (2020); Nam et al. (2017); Netzer et al. (2012); Pan et al. (2019); Piezunka and Dahlander (2015, 2019); von Hippel and Kaulartz (2021); Wang et al. (2021); Zhou et al. (2018)
<i>Information theory</i>	
Entropy	Glasserman and Mamaysky (2019); Liu et al. (2020)
Perplexity	Liu et al. (2020)
Kullback-Leibler divergence	Cookson and Niessner (2020); Liu et al. (2020)
Continued on next page	

Table 2.3 – continued from previous page

Measures of Document Vectors Similarity	References
Jensen–Shannon divergence	Corritore et al. (2020)
Pointwise Mutual Information / Lift	Archak et al. (2011); Kim et al. (2015); Netzer et al. (2012)
<i>Similarity measure from a third party</i>	
RavenPack’s relevance score	Di and Laux (2021)
<i>Similarity of one document to a group of documents</i>	
Difference between the individual and prototypical (e.g. industry average) document vectors; Centroid	Liu et al. (2020); Taeuscher et al. (2021)
Silhouette value	Bao and Datta (2014); Theurer et al. (2021)

Summary of findings on text similarity measures

Figure 2.5 is a heatmap, mapping management constructs (row titles) and common text similarity measures (column titles). The number represents the frequency of papers adopting a certain text similarity measure to represent a certain management construct, and the color becomes darker when the number of papers is larger. As is discussed at the beginning of Section 2.2.1, the management constructs are divided into two types, those directly measured by text similarity against a *benchmark* (summarized in Table C.1), and cross-sectional or temporal ones measured by text similarity against *other documents* in the corpus (summarized in Table C.2), e.g., convergence of opinions.

The cross-sectional or temporal measures, listed in the second panel of Figure 2.5, use more types of methods than those directly measured by text similarity as is listed in the first panel. This is because the constructs directly measured by text similarity is constrained by the benchmarks. In addition to that, no obvious pattern can be found in the mapping between management constructs and common text similarity measures. This implies that no similarity measure dominates others, and they complement each other.

The most popular measure is cosine similarity, with 27 out of 67 papers adopting it. The second popular type of measures are those belonging to information theory, e.g., entropy, Kullback-Leibler divergence, etc. The popularity of the two measure does not imply that they are better than other measures, and all of them have weaknesses. For example, cosine

similarity (based on Euclidean distance) is not reliable when the size of text vectors is too large, because Euclidean distance is based on L2 norm and does not work well with high-dimensional data (Domingos, 2012).

Considering that no text similarity measure is perfect, a plausible suggestion is using more similarity measures in one paper as robustness tests. That is actually what management researchers do over the past few years. Figure 2.6 shows the yearly trend of the average number of text similarity measures used in each paper in my sample. Before 2017, all the papers only adopt one measure, while after 2018, there is a trend that increasing number of methods are used. For example, Liu et al. (2020) use seven different methods to measure text similarity, including cosine similarity, Kullback-Leibler divergence, number of words overlapping, etc.

		Overlapping terms	Vector distance / angle	Information theory	Third party score	Individual against a group
Management constructs directly measured by text similarity	Innovation, novelty	1	4			
	Customer agility		1			
	Quality of idea		1			
	Intellectual distance / experience relevance	1	3			
	Female-typedness		1			
	Normative legitimacy	1				
	Market structure	1	3	1		
	Unusualness	1		1		
Cross-sectional or temporal management constructs measured by text similarity	Convergence / divergence of opinions / ideas, agreement /disagreement		4	2		
	Match / relevance of content		6	2		1
	Personality similarity		1	1		
	Product uniqueness / similarity		3			
	Culture congruence / heterogeneity			1		
	Industry convergence	1		1		
	Similarity of brands' social tags		1			
	Coherence/changes of identity		1	1		
	Bibliographic coupling network		1			
	Relationship between risk types					1
	Relationship of the firm and the underlying news story				1	
	Distinctiveness between entrepreneurial story and the prototypical					1
	Distinctiveness of employer image or with third-party employer ratings		2			1
	The effect of social media advertising content on customer engagement	1				

Fig. 2.5 Mapping between management constructs and common text similarity measures. The number represents the frequency of papers adopting a certain text similarity measure (column titles) to represent a certain management construct (row titles). The color becomes darker when the number of papers increases. Please note that one paper may adopt more than one text similarity measures.

Text similarity can be further used in clustering (Bao and Datta, 2014; Timoshenko and Hauser, 2019) and network analysis (Chen et al., 2021, 2019; Hoberg and Phillips, 2016;

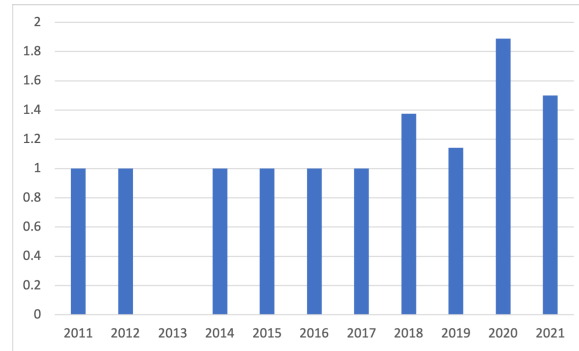


Fig. 2.6 Yearly trend of the average number of similarity methods used in each paper. There is no papers using text similarity in 2010 and 2013.

Netzer et al., 2012), as the similarity or distance reflect the relationship between different pairs of observations.

2.3 Text similarity measures at different levels

As discussed previously, a requisite step to measure text similarity is text vectorization (Section 2.2.2), which is then used to construct a similarity measure (Section 2.2.3). While in this section, I employ alternative document vectorizations, and focus on the calculation of text similarity at different aggregation levels.

Based on the literature review on text similarity, text similarity can be divided into three levels: **micro-micro** (similarity between individual documents), **micro-macro** (similarity between a document and a class of documents), and **macro-macro** (similarity between classes of documents). At all three levels, one can consider the similarity both within the same class and across different classes of documents. Therefore, I have 3×2 different text similarity measures, as are summarised by Table 2.4.

In the literature I reviewed in Section 2.2, most measures of text similarity are at the micro-micro level, for which standard measures of vector distance/angle can be employed, while very few are at the other two levels. I will therefore focus on the less explored micro-macro and macro-macro levels. In this section, I will introduce all the within- and across-class text similarity measures corresponding to the last two rows of Table 2.4. I use similarity rather than dissimilarity to ensure consistency across all the measures. In the sequel, I will use the terms “similarity” and “convergence” interchangeably.

	Across-class	Within-class
Micro-micro	Text similarity between two individual documents, which belong to different classes.	Text similarity between two individual documents, which belong to the same class.
Micro-macro	Text similarity of one document to a whole class of documents, and the document does not belong to the class.	Text similarity of one document to a whole class of documents, and the document belongs to the class.
Macro-macro	Text similarity among different classes of documents.	The level of concentration of one class of documents, which is the overall evaluation of how similar the documents in the class are to each other.

Table 2.4 Definition of similarity at different aggregation levels.

2.3.1 Micro-macro level convergence

Micro-macro across-class convergence

I use *negative silhouette value* to measure the convergence of one individual document to a class of documents, and the individual document does not belong to the class. This measure will be applied in Chapter 4 as a special case when there are only two classes. The silhouette value is a measure used in unsupervised learning, reflecting how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each observation. A low value indicates that the object is poorly matched to its own cluster and well matched to neighboring clusters (Rousseeuw, 1987). Silhouette value can also be applied to labelled data to measure the across-class and within-class distance, where the clusters are replaced by classes.

Suppose I have M different classes, $M \geq 2$. Define $C = \{c_1, \dots, c_M\}$, which is the set of all classes. For a given $1 \leq k \leq M$, the given class c_k contains N_k document vectors, i.e. $c_k = \{y_{k1}, y_{k2}, \dots, y_{kN_k}\}$, where $y_{k1}, y_{k2}, \dots, y_{kN_k} \in \mathbb{R}^{1 \times p}$ are individual document vectors and p is the number of features in each document vector.

To calculate convergence for each document vector $y_{ki} \in c_k$ ($1 \leq i \leq N_k$) in Class k , I define the quantities:

$$a_{ki} = \frac{1}{N_k - 1} \sum_{y_{kj} \in c_k, j \neq i} d(y_{ki}, y_{kj}),$$

$$b_{ki} = \min_{c_l \in C, l \neq k} \frac{1}{N_l} \sum_{y_{lj} \in c_l} d(y_{ki}, y_{lj}),$$

where d represents the Euclidean distance between document vectors. Here, a_{ki} represents the average distance / dissimilarity of the document, y_{ki} , to other documents within the same class (cohesion), while b_{ki} reflects the average distance / dissimilarity to all the documents in the nearest class (separation). Then, my measure of across-class convergence of a document y_{kj} is given by the *negative silhouette value* defined as

$$nsv_{ki} = -\frac{b_{ki} - a_{ki}}{\max\{a_{ki}, b_{ki}\}}. \quad (2.1)$$

The original silhouette value reflects the scaled *distance* of one observation to the nearest class. The negative sign, used here, reflects the extent to which a document is *converged* to documents in different classes,

Micro-macro within-class convergence

Now I measure the convergence of one document to all the documents in the class it belongs to, e.g., how similar a document is to its own industry. I use 'part' of the silhouette value, a_{ki} , the cohesion measure, which is the average distance of one document to all the other documents in the same class. I use the negative value to turn distance into convergence. It is scaled by the maximum of all the pairwise distance, because some classes are more scattered, and those even at the center (the most similar to its own class, or has the highest within-class convergence) can have bigger average distance compared to less scattered classes. Therefore, the within-class convergence is measured by Equation (2.2).

$$na_{sv_{ki}} = \frac{-a_{ki}}{\max_{y_{ki}, y_{kj} \in c_k} \{d(y_{ki}, y_{kj})\}}. \quad (2.2)$$

2.3.2 Macro-macro level convergence

Macro-macro across-class convergence

To build the macro-macro convergence across industry measure, I borrow the idea from Fisher Discriminant Ratio (FD Ratio) in Fisher Discriminant Analysis (FDA) (Fisher, 1936; Fukunaga, 2013). FDA projects the data to a line such that the two classes are well-separated. When doing the projection, FDA maximise the FD Ratio of between-class scatter and within-class scatter so that the projected data are easy to be separated for classification purpose. Instead of doing classification and using the projected data, I use the original data to calculate the FD Ratio, which is a good measure of class separation. I generalize the definition to more than two classes, so the between-class definitions are replaced by the across-class.

The within- and across-class covariance are give by S_W and S_A respectively:

$$S_W = \sum_{c_k \in C} \left[\frac{1}{N_k} \sum_{y_{ki} \in c_k} (y_{ki} - \mu_k)^T (y_{ki} - \mu_k) \right] \in \mathbb{R}^{p \times p}, \quad (2.3)$$

$$S_A = \sum_{k \neq l, c_k, c_l \in C} (\mu_k - \mu_l)^T (\mu_k - \mu_l) \in \mathbb{R}^{p \times p}, \quad (2.4)$$

where μ_k is the average vector of all the standardized document vectors in c_k . S_W is the sum of covariance matrix of all classes, reflecting within-class spread. S_A is the across-class covariance, reflecting the difference among pairwise mean vectors of different classes at all dimensions, i.e. across-class separation.

The traditional FD Ratio of multi-classes is defined in Equation (2.5):

$$FD\ Ratio = \frac{\det(S_A)}{\det(S_W)}. \quad (2.5)$$

The FD Ratio reflects the separation among classes, scaled by the within-class separation. Determinant of a matrix is a single number to summarize the volume of the matrix (Margalit et al., 2017). Borrowed from the idea of silhouette value, I construct the macro-macro across-class convergence as in Equation (2.6). Again, the across-class distance, reflected by $\det(S_A)$, is scaled in the same way as Equation (2.1). Since I measure convergence rather than distance, I have a negative sign. Therefore, the macro-macro across-industry convergence metric is measured by *negative standardized FD Ratio* (*nsfdr* in short):

$$nsfdr = -\frac{\det(S_A) - \det(S_W)}{\max\{\det(S_A), \det(S_W)\}}. \quad (2.6)$$

The value of *nsfdr* is standardized to -1 to 1 . A value close to 1 implies high convergence, while -1 implies high divergence.

Macro-macro within-class convergence

To calculate the macro-macro within-class convergence, the first step is standardization of all features by class, because I calculate covariance matrix later, which may be biased by extreme values. Denote the matrix

$$Y_k = [y_{k1}^T, y_{k2}^T, \dots, y_{kN_k}^T]^T = \begin{bmatrix} f_{k11} & \dots & f_{k1p} \\ \dots & \dots & \dots \\ f_{kN_k1} & \dots & f_{kN_kp} \end{bmatrix} \in \mathbb{R}^{N_k \times p} \quad (2.7)$$

representing all the documents belonging to the same Class c_k . f_{kij} , $i \in \{0, \dots, N_k\}$, $j \in \{0, \dots, p\}$ is the j th feature of the i th document in Class k . I standardise each feature across all the documents, and denote the standardised matrix as

$$\tilde{Y}_k = [\tilde{y}_{k1}^T, \tilde{y}_{k2}^T, \dots, \tilde{y}_{kN_k}^T]^T = \begin{bmatrix} \tilde{f}_{k11} & \dots & \tilde{f}_{k1p} \\ \dots & \dots & \dots \\ \tilde{f}_{kN_k1} & \dots & \tilde{f}_{kN_kp} \end{bmatrix} \in \mathbb{R}^{N_k \times p} \quad (2.8)$$

where

$$\begin{aligned} \tilde{f}_{kij} &= \frac{f_{kij} - \mu_{kj}}{\sigma_{kj}}, \\ \mu_{kj} &= \frac{1}{N_k} \sum_{s=1}^{N_k} f_{ksj}, \\ \sigma_{kj} &= \sqrt{\frac{1}{N_k} \sum_{s=1}^{N_k} (f_{ksj} - \mu_{kj})^2}. \end{aligned}$$

I use the negative determinant of the covariance of the text vectors to measure the macro-macro within-industry convergence. The determinant of the covariance matrix reflect how spread the data are (Ko et al., 1995). Hence, as a convergence measure, I use the negative of the value to measure how concentrated all the individuals in a class are.

The macro-macro within-class convergence is defined as follows:

$$det_cov = -\ln[\det(\text{Cov}(\tilde{Y}_k))], \quad (2.9)$$

where $\tilde{Y}_k$ is defined in Equation (2.8). The logarithm in the deviation is there to allow for variations in the scale of $\det(\text{Cov}(\tilde{Y}_k))$.

All the text convergence measures at different levels are summarized in Table 2.5.

	Across-class	Within-class
Micro-micro	Cosine similarity	Cosine similarity
Micro-macro	Negative silhouette value, ns_v in Equation (2.1)	Negative average distance of one firm to all the other firms in the same industry, scaled by the maximum of all the pairwise distance, na_{sv} in Equation (2.2)
Macro-macro	Inverse of the ratio of across-class scatter and within-class scatter, $nsfdr$ in Equation (2.6)	Negative determinant of the covariance of the text vectors, det_cov in Equation (2.9)

Table 2.5 Convergence measures based on text similarity at different levels.

2.4 Application of text similarity measures for quantifying industry convergence

To illustrate the application of the convergence measures based on text similarity that I introduced above, I apply them to Section 1A Risk Factors in 10-K financial reports, and focus on the insurance and banking industries as the two classes. U.S. 10-K forms, downloaded from Filings Expert, are annual reports required by the U.S. Securities and Exchange Commission (SEC), giving a comprehensive summary of a company's financial performance¹. The Risk Factors section includes information about the most significant risks that apply to the company or to its assets. My data cover all the 221 insurers and 934 banks that submitted 10-K reports to U.S. Securities and Exchange Commission (SEC) from 2006 to 2018. Companies are categorized to different industries according to their NAICS codes from the Orbis database. To ensure the industry classification is correct, I also manually read each company's profile on Bloomberg (Bloomberg, 2021) to check the business that they primarily run.

I carry out the analysis for each year separately, so that I can check for any trends across the thirteen years in my sample. I apply convergence measures to four document vectorizations: term frequency (TF), term frequency-inverse document frequency (TFIDF),

¹Here is a sample 10-K form of American International Group, Inc.:
<https://www.sec.gov/Archives/edgar/data/5272/000095012309003734/y74794e10vk.htm>

Doc2vec (D2V) and Latent Dirichlet Allocation (LDA, definitions of these four methods can be found in Section 2.2.2) and assess the extent to which results depend on the way that documents are summarized as vectors. My Doc2vec model has 100 dimensions, i.e. 100 features in the text vector, trained on the same 10-K data. Given the specialised vocabulary used, it is not trained on external dataset because there are many unique items in my dataset that are not included in other external ones. One of the key parameters, window (the maximum distance between the current and predicted word within a sentence), is equal to 8 after I carried out cross validation. The topic number of LDA model is equal to 6, chosen according to interpretability, and so the corresponding text vector has 6 features.

2.4.1 Micro-macro level convergence

Yearly trend of micro-macro convergence

The yearly trend of the micro-macro across-industry convergence, measured by negative silhouette values (nsv_{ki}) in Equation (2.1), can be found in Figure 2.7, and the within-industry convergence (na_sv_{ki}) in Equation (2.2) can be found in Figure 2.8. I show the means and 95% confidence intervals of the measures in each year by industry.

As seen in Figure 2.7, at the micro-macro level, insurers have higher across-industry convergence than banks, implying that insurers' risk disclosure is more similar to banks, while banks' risk disclosure is more industry-idiosyncratic. This conclusion is consistent for all vectorizations. Different from symmetric measures at the micro-micro level, e.g. cosine similarity, the similarity at micro-macro level is asymmetric between insurance and banking industries. They reveal a different feature of the data compared to those at the micro-micro level.

In Figure 2.8, I show the micro-macro within industry convergence as in Equation (2.2). Here the converse is observed: for all vectorisations, banks have higher with-industry convergence, confirming the previous observation that they tend to be more idiosyncratic than insurers.

The last observation motivates the question of how, for a given firm, its convergence to another industry ('across') correlates with the convergence to its own industry ('within'). In Figure 2.9, I plot these convergence measures for all firm-years by industry, after standardizing both measures by industry. The relationship is generally negative except for insurers when using TF and Doc2Vec vectorizations. For all vectorizations the Pearson's and Kendall's correlation coefficients are significant at 1%. For the banking industry, the negative correlation echoes the previous observations around Figures 2.7 and 2.8. This confirms the intuition that if a bank's document is similar to a prototypical document representing the

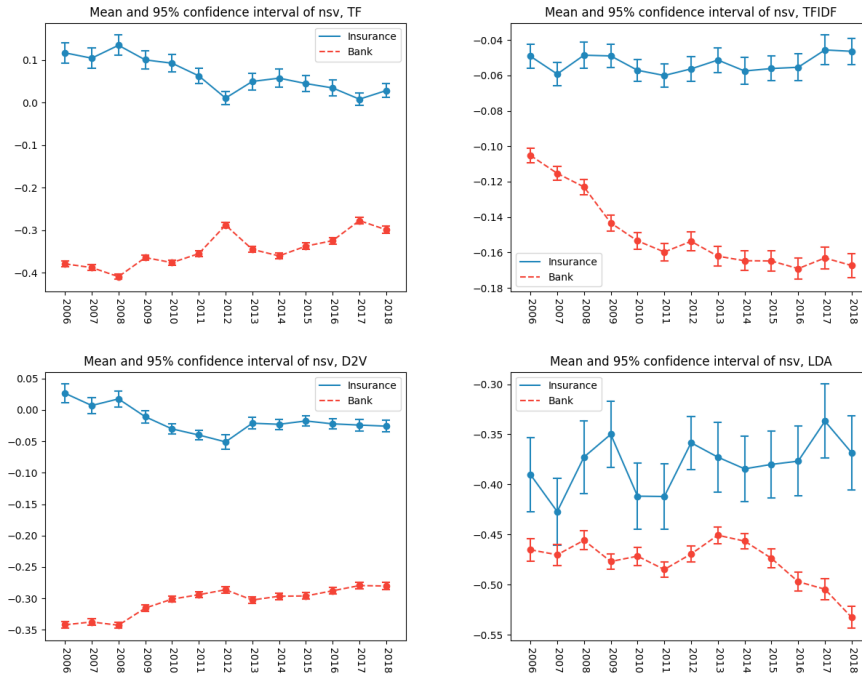


Fig. 2.7 The yearly means and 95% confidence intervals of the micro-macro across-industry convergence metrics.

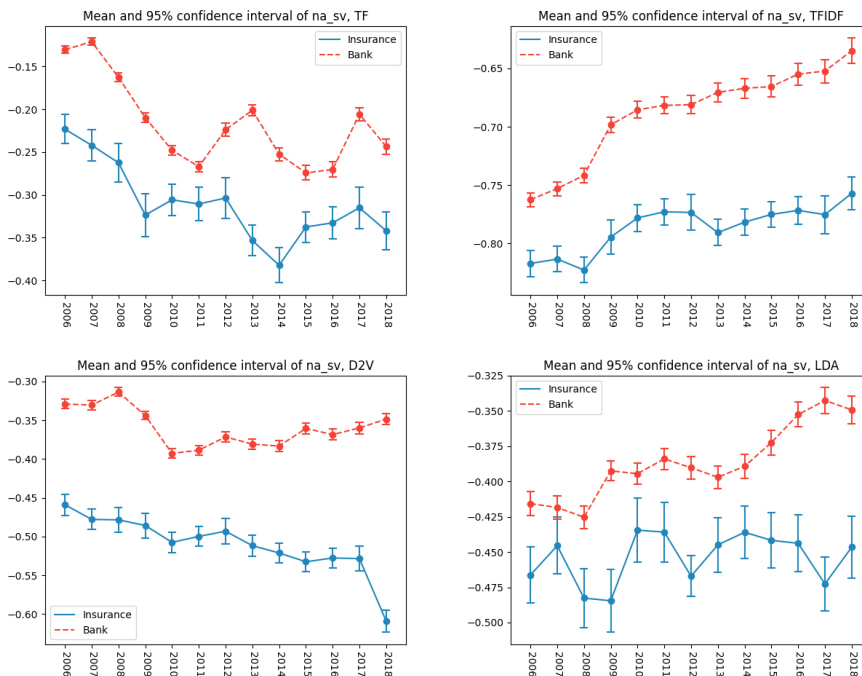


Fig. 2.8 The yearly means and 95% confidence intervals of the micro-macro within-industry convergence metrics.

banking industry, it will be dissimilar to the insurance industry, and vice versa. However, I can not conclude so for insurers, as the results are not robust.

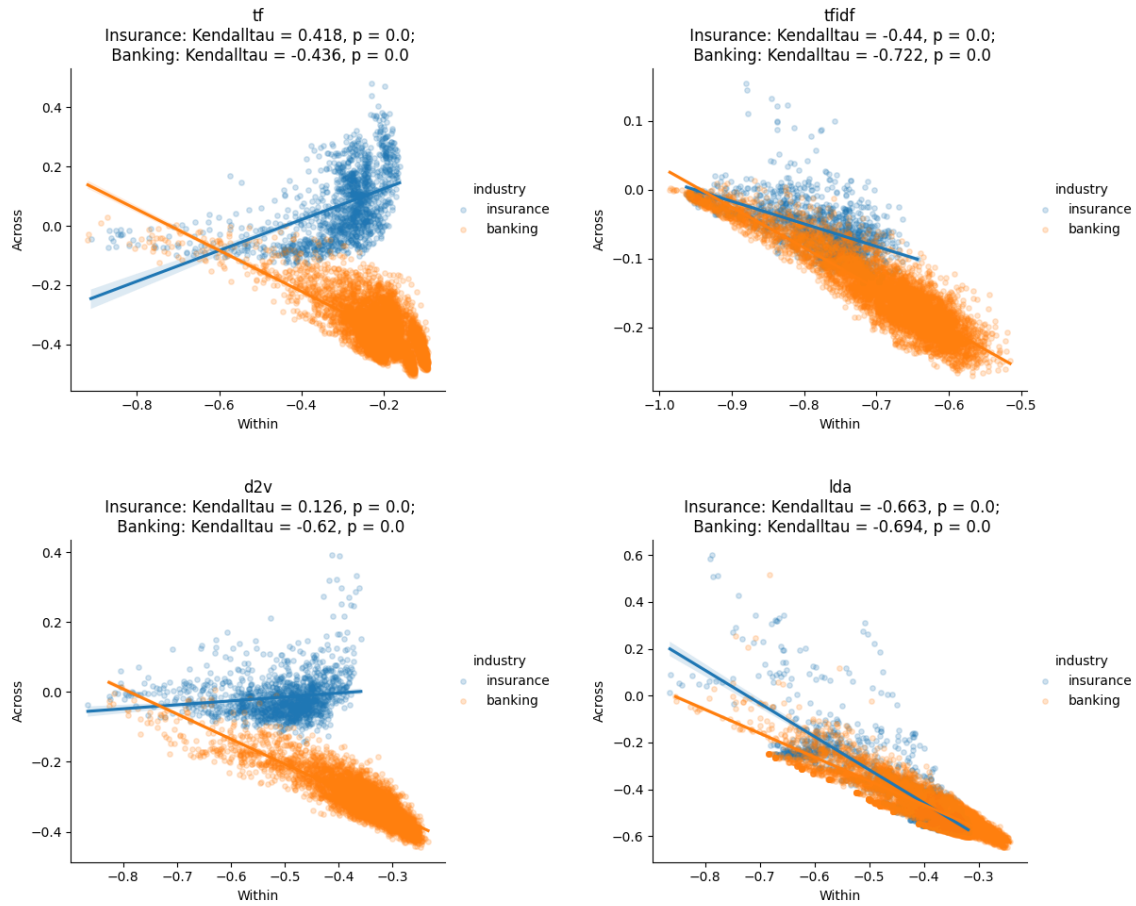


Fig. 2.9 The relationship of the across-industry convergence against the within-industry convergence at the micro-macro level.

The convergence measured are standardised by industry. The blue and Orange scatters represent insurance and banking industry respectively. Kendall tau coefficients of both industries and their p-values are reported at the top of each sub-figure.

Consistency across the four text vectorizations

Following the previous analysis, a natural question then is: to what extent are the measures based on the four text vectorization methods more broadly consistent?

Figure 2.10 shows the correlation matrix of the micro-macro *across-industry* convergence of the four vectorizations. All the variables are standardized by industry. All the pairwise Pearson and Kendall correlation coefficients are significantly positive, implying that the

measures using the four vectorizations are, at least to an extent, consistent. I report Kendall correlation coefficients in the figure because of the non-linear pattern in the data shown in Figure 2.10, which makes Kendall correlation more reliable (Kendall, 1938). LDA and TF capture relatively different features of the data, i.e. topics distribution and simple term frequency, and thus have the lowest correlation. Interestingly, TF and Doc2vec are the most consistent, so they capture similar features. I will discuss the four vectorization methods in Section 2.5. Figure 2.11 shows the correlation matrix of the micro-macro *within-industry* convergence. All the pairwise Pearson and Kendall correlation coefficients are significantly positive except the one between TF and TFIDF. That is because TFIDF adjust for the fact that some words appear more frequently in general, rather than use simple term frequency (Rajaraman and Ullman, 2011). The same as the relationship in across-industry, convergence measures of TF and Doc2vec are the most consistent.

However, when it comes to the yearly trend, the four text vectorizations are inconsistent, as seen in Figures 2.7 and 2.8. For example, in Figure 2.7 and if I focus on the banking industry, the convergence of TF and Doc2vec shows a generally increasing trend, while that the trend of TFIDF and LDA generally decreasing. I conclude that I cannot robustly state that there is an increasing or a decreasing yearly trend in the micro-macro convergence from my data, for either industry.

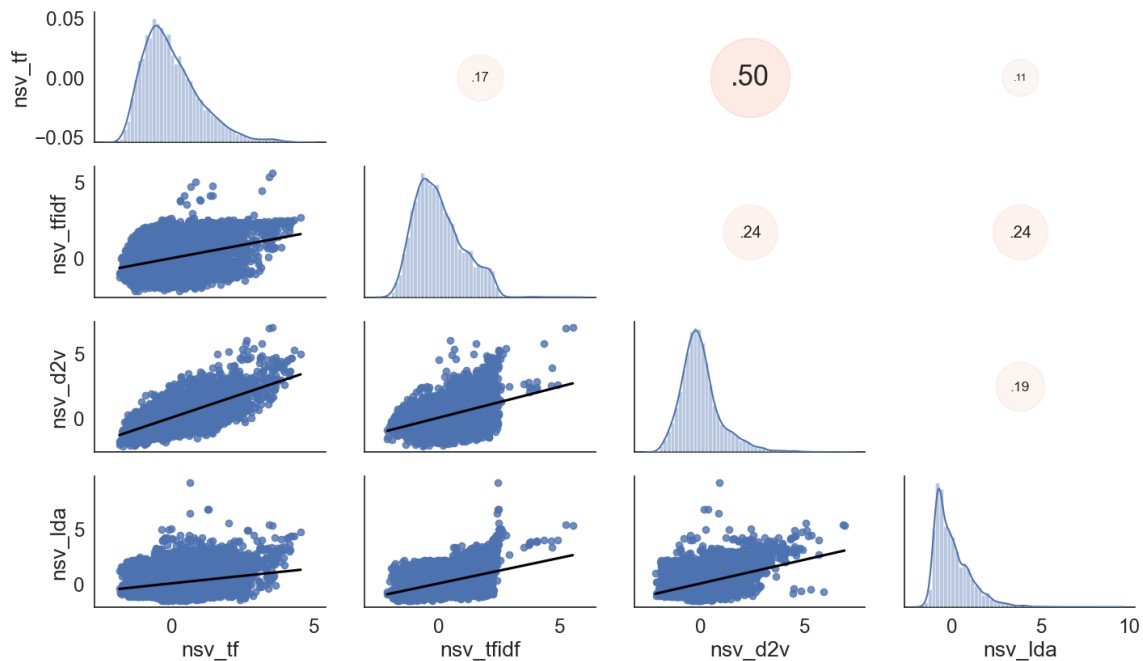


Fig. 2.10 The correlation matrix of the micro-macro across-industry convergence of the four vectorizations.

Standardized by industry. The upper triangle reports all the Kendall correlation coefficients.

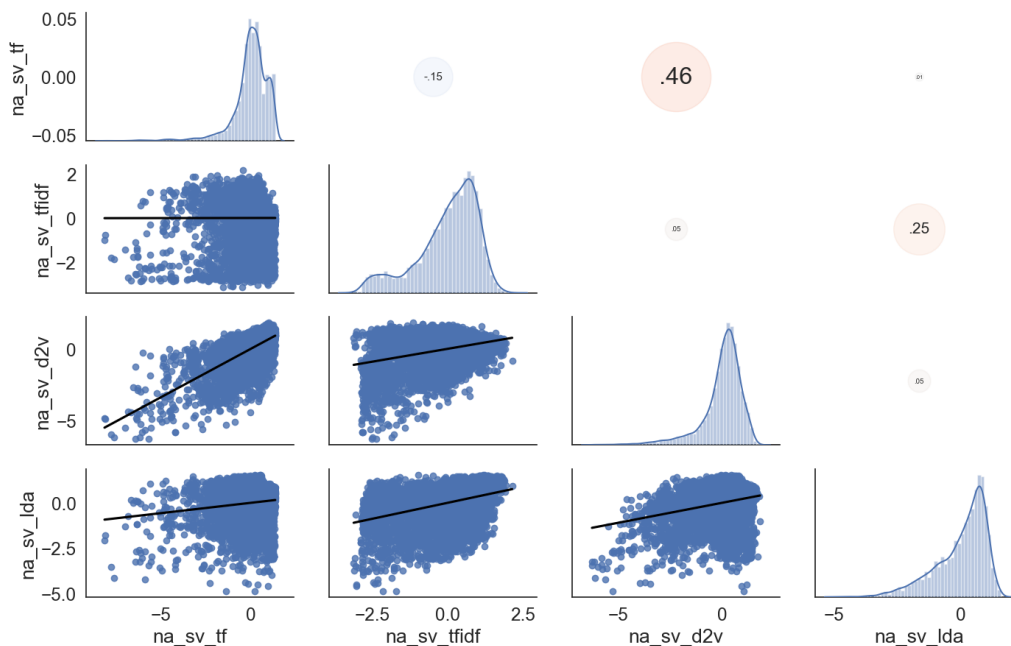


Fig. 2.11 The correlation matrix of the micro-macro within-industry convergence of the four vectorizations.

Standardized by industry. The upper triangle reports all the Kendall correlation coefficients.

Micro-macro convergence and prediction of systemic risk

How can these text-based convergence measures be used? One application is to predict systemic risk. As I mentioned in Section 4.2, text data reflect the attention or cognition of managers, which guides human behaviour, and thus can be considered as an antecedent of systemic risk. Here, I examine the extent to which my document convergence measures can be used to predict systemic risk.

I evaluate systemic risk through daily stock return correlations, during the 2020. This year corresponds to the break out of the Covid-19 Pandemic, which is a systemic risk event. Convergence in management attention in earlier years may cause firms to react similarly to an external shock, e.g., the Pandemic. Here I measure systemic risk using the second category discussed in Section 1.2.3, i.e. systemic risk measures using conditional stress: I try to reflect an individual firm's correlation conditional on the external shock, Covid-19 Pandemic.

I use simple systemic risk measures, applying both across- and within-industry. The across-industry systemic risk, denoted as $\tau_{oppo20_{ki}}$, is defined as the Kendall correlation coefficient between an individual insurer's (resp. bank's) daily stock return and the return on the S&P 500 Bank Index (resp. Insurance Index) during the Year 2020. This is similar to the $StockCorr_{it}$ in Chapter 4. The within-industry systemic risk, denoted as τ_{own20} , is

defined as the Kendall's Tau between the daily return of the firm stock, and the daily return of the index of its *own* industry during the Year 2020.

I then consider how my micro-macro convergence measures in 2018 are correlated with systemic risk measures during the Pandemic for the same firms. I have only 64 insurers and 114 banks in the sample because there are missing values in the stock data and only one year is considered. Year 2018 is chosen because it is the last year in our dataset. In future work, I may use data of other years, or other data sources, to check whether management attention in earlier years, reflected by text data, can be used to predict systemic risk.

Figure 2.12 shows the scatter plots of systemic risk ($\tau_{oppo20_{ki}}$) during the Covid-19 Pandemic against document convergence (nsv_{ki}) in 2018, at the **across-industry** level. The left column is insurance, while the right column represents the banking industry. Each row shows the plots of one of the four vectorizations. 1% outliers are excluded. I can see the two variables are significantly positively correlated in the banking industry according to the Kendall correlation coefficients and their p-values. For the insurance industry I observe no significant correlation, possibly because the number of insurers in the sample is only 64 and is not big enough. Therefore, if a bank's risk disclosure was converged to the insurance industry and not industry-idiosyncratic in 2018, it would experience high across-industry systemic risk (i.e. its stock return would be more correlated to the other industry) during the Covid-19 Pandemic.

Similar to Figure 2.12, Figure 2.13 shows the scatter plots of systemic risk ($\tau_{own20_{ki}}$) during the Covid-19 Pandemic against document convergence ($na_{sv_{ki}}$) in 2018, at the **within-industry** level. For the banking industry, the relationship is negative, and it is also generally negative for the insurance industry. The negative correlation implies that if a firm's risk disclosure was similar to a typical industry document in 2018, it would have lower within-industry systemic risk during the Pandemic.

The results imply that paying attention to industry-specific risk helps reduce systemic risk within the industry (Figure 2.13), because problems within the industry can be found timely and resources are allocated accordingly to solve them. However, the impact of attention to the other industry is ambivalent, similar to what is discussed in the conclusion of Chapter 4. On the one hand, inter-industry and economy-wide problems can be detected more timely because attention is paid to them. On the other hand, attention to the other industry connects different industries, exposing them to across-industry systemic risk. Results in Figure 2.12 show that the second effect dominates in my data and research context. In summary, management attention to *risk* itself (both within- and across-industry) helps reduce risk, but attention to the *other industries* connects different industries and expose firms to across-industry systemic risk.

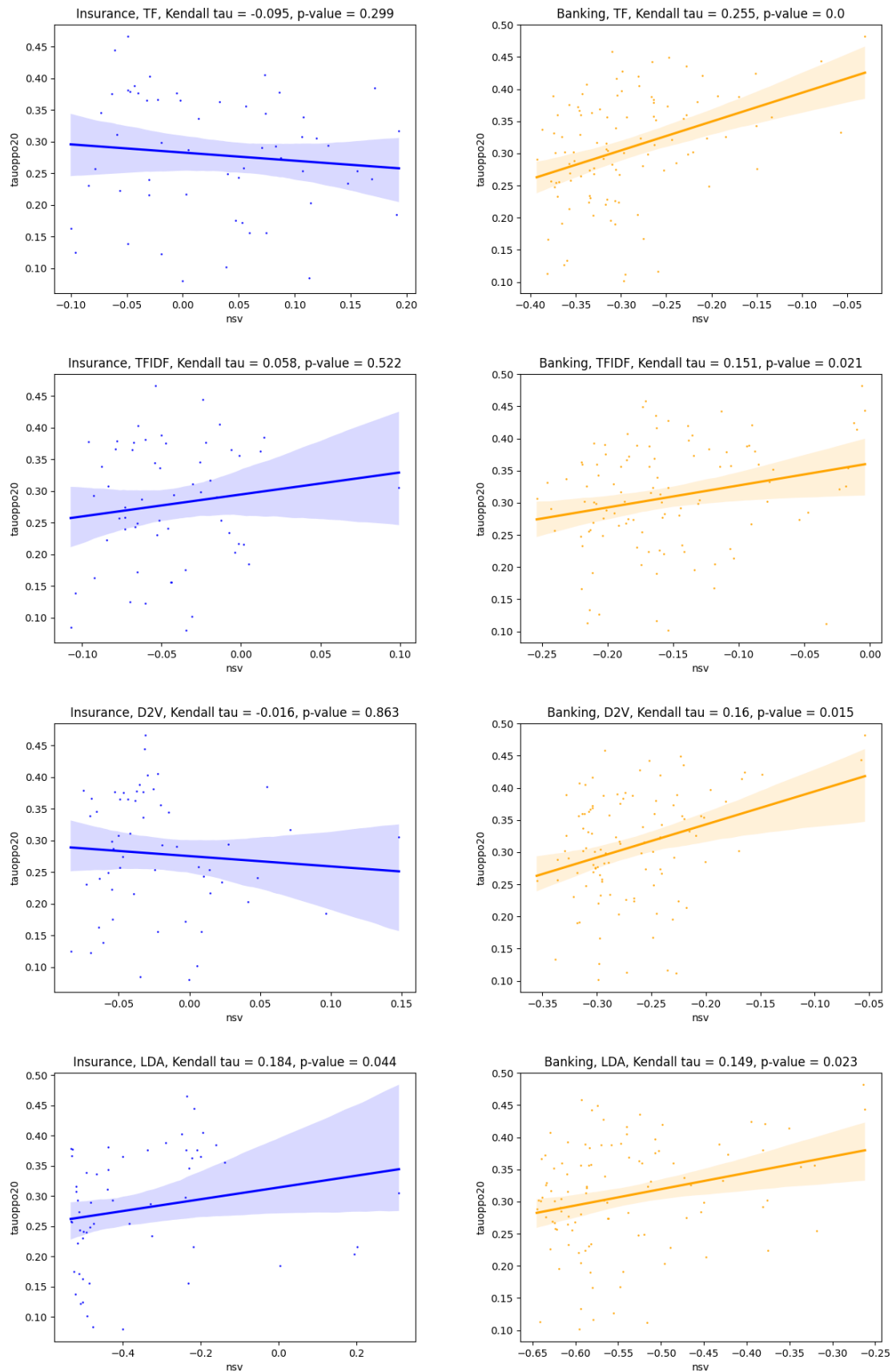


Fig. 2.12 Relationship between systemic risk ($\tau_{oppo20_{ki}}$) during the Covid-19 Pandemic and document convergence (nsv_{ki}) in 2018, at the across-industry level. Kendall correlation coefficients, together with the p-values, are reported at the top of each sub-figures.

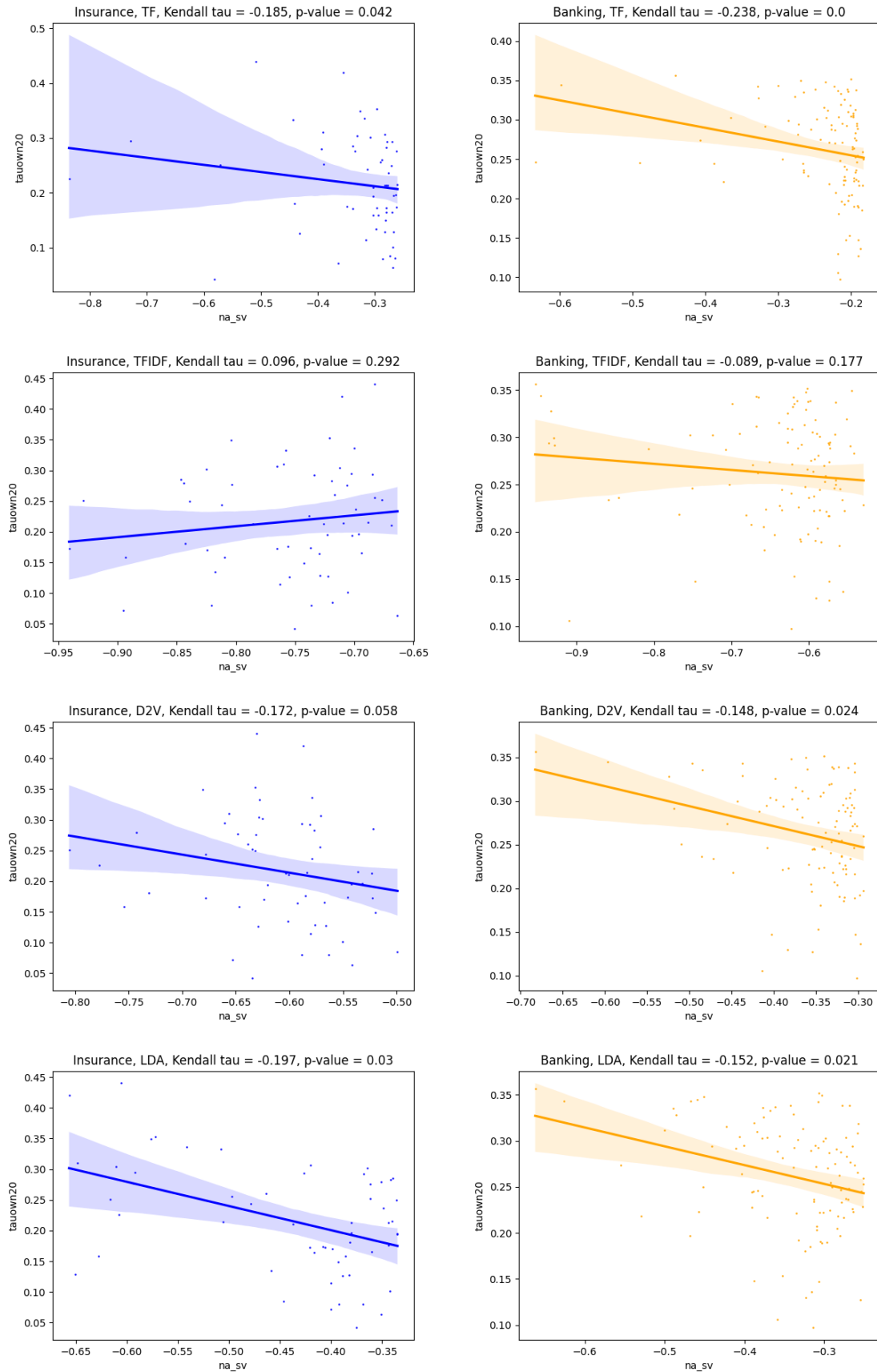


Fig. 2.13 Relationship between systemic risk ($\tau_{own20_{ki}}$) during the Covid-19 Pandemic against document convergence ($na_{sv_{ki}}$) in 2018, at the within-industry level. Kendall correlation coefficients, together with the p-values, are reported at the top of each sub-figures.

2.4.2 Macro-macro level convergence

Now, instead of studying the convergence at the individual firm against an industry, i.e. micro-macro, level, I turn my attention to industry (macro-macro) level, using the convergence measures in Equation (2.6) (*nsfdr*, across-industry) and (2.9) (*det_cov*, within-industry).

For TF and TFIDF vectorization, the feature dimension is 18754, much larger than the total observation numbers, 7565. This cause inaccuracy in the estimation. For LDA vectorization, the features in topic distribution vector are highly correlated. Instead of using the text vectors directly, I first apply Principle Component Analysis (PCA) to the original document vectors by industry in each year separately. For LDA vectorization, the number of components are 5. For the other three vectorizations, the number of components equal to 77, because the smallest observation number of insurance in one year is 78.

Yearly trend of macro-macro convergence

When calculating the macro-macro across-industry convergence, *nsfdr* in Equation (2.6), I only use the first principle component of the original data, because the rank of S_A is 1 and will get a zero $\det(S_A)$ if I involve more than the one principle components. Choosing the first principle component is consistent with common practice when applying FDA (Fisher, 1936; Fukunaga, 2013).

The yearly trend of macro-macro across- and within-industry convergence can be found in Figure 2.14 and Figure 2.15 respectively.

According to Figure 2.14, at the macro-macro across-industry level, the convergence of LDA provides the clearest trend, which peaked around 2008 and 2012 and is consistent with the history of the Subprime mortgage crisis and the European sovereign debt crisis. The good interpretability of LDA in this figure may be related to the fact that LDA reveals the semantic meaning of the corpus, and involve the most amount of human interpretation (e.g., when choosing the best number of topics, I manually read all the topics and choose the number that produces the most interpretable topics). The convergence of TF has the highest value close to 1, implying high convergence between the insurance and banking industry, while those of TFIDF and Doc2vec are close to -1 and relatively stable, implying constant high divergence between the two industries.

When calculating the macro-macro within-industry convergence, *det_cov* in Equation (2.9), for LDA and Doc2vec vectorizations, I choose the first several Principle Components (PCs) that explain more than 90% of variance to replace the text vectors; for TF and TFIDF, the explain ratio is only 30% because 77 is much smaller than 18754.

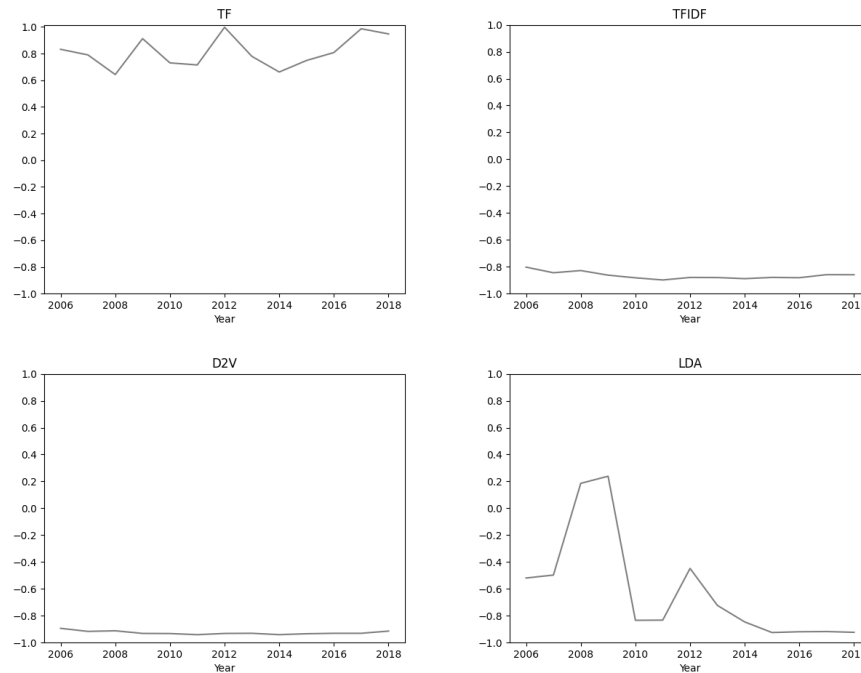


Fig. 2.14 Macro-macro across-industry convergence using the first principle component of the original text vectors.

In Figure 2.15, regarding the within-industry convergence, the banking industry is generally decreasingly concentrated, while the within-industry convergence of insurance fluctuates. This conclusion is consistent across all the four vectorizations.

2.5 Choice of text vectorization methods

I have four text vectorization methods, TF, TFIDF, Doc2vec, LDA. Though the measures at the micro-macro levels using four vectorizations are generally positively correlated (Figure 2.10 and 2.11) and some results are consistent, other results are quite inconsistent, especially regarding the yearly trend (Figure 2.7, 2.8, 2.14, 2.15). I conclude that different vectorizations extract very different types of information from data. According to Figure 2.10 and 2.11, TF and Doc2vec extract similar information from my data.

To decide which vectorization to use in practice, I need a deeper examination of whether the features of the text data that the vectorization extracts are well matched with the real-world problem that the researcher is trying to address. LDA is very effective in reflecting broad topics, but may under-represent niche content with low frequency, especially if the number of topics is very small. On the contrary, when using TFIDF, term frequency is offset

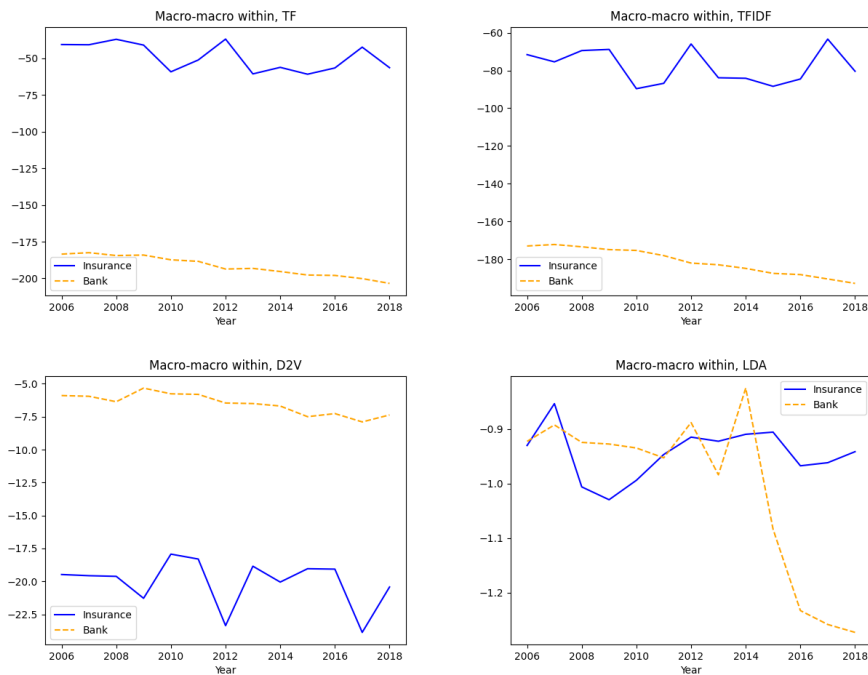


Fig. 2.15 Macro-macro within-industry convergence using the negative determinant of the covariance of the text vectors.

by the number of documents that contain the word, so more weight is put on rarely used terms.

Nonetheless, using purely technical methods, I can carry out a partial validation of the different vectorizations. I do classification using the four vectorizations, and use common classification evaluation metrics, e.g., accuracy rate, to evaluate the performance of each vectorization. In my case, the labels of the two classes are insurers and banks. Good classification performance of a vectorization method, i.e., classifying observations correctly to the right industry, implies a good representation of features of the data sample.

I use logistic regression as the classifier, and cross validate with 20 folds for each year separately. As I have far more banks than insurers in my sample, I make sure all folds are stratified and re-balanced, to have the same proportion of insurers and banks in them. I run each classification with different vectorizations and report the performance results, averaged over all folds.

My classification evaluation metrics include:

- (a) Accuracy rate, which is the number of correct predictions divided by the total number of predictions. Higher values of accuracy rates indicate better classification performance.

- (b) F1 score is also a measure of accuracy applied specifically to binary classification (Powers, 2020). It is the harmonic mean of precision and recall. If I define insurers as the positive class, then *precision* is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of observations labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are observations incorrectly labelled as belonging to the class); *recall* is the number of true positives divided by the total number of observations that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are observations that were not labelled as belonging to the positive class but should have been). The definition of F1 score is given below, where TP, TN, FP, FN represent true positive, true negative, false positive and false negative respectively. Higher value of F1 scores indicate better classification performance.

$$F1 = \frac{2}{recall^{-1} + precision^{-1}} \quad (2.10)$$

$$= \frac{2TP}{2TP + FP + FN} \quad (2.11)$$

- (c) Cohen's kappa score, is generally thought to be a more robust measure than simple percent agreement calculation, as it takes into account the possibility of the agreement occurring by chance (Banerjee et al., 1999). Its definition is given below. Higher value of kappa scores indicate better classification performance.

$$\kappa = \frac{2(TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \quad (2.12)$$

- (d) The average of *Blurring*, which is defined as the prediction probability of the wrong type (*Blurring* will be formally defined in Section 3.3.2). Higher value of it implies that the classifier using a certain vectorization is less confident to classify firms to the right class compared to other vectorizations, and thus has worse performance. To achieve consistency with the other three performance evaluation measures, I use one minus average *Blurring* of the year as the fourth classification performance measure.

Results of the classification evaluation metrics in each year are reported in Figure 2.16. In my dataset, all the four vectorizations have good performance. Doc2vec performs the worst, especially for later years. This is because Doc2vec is trained on my own data, which is very limited. All the other three vectorizations have close performance.

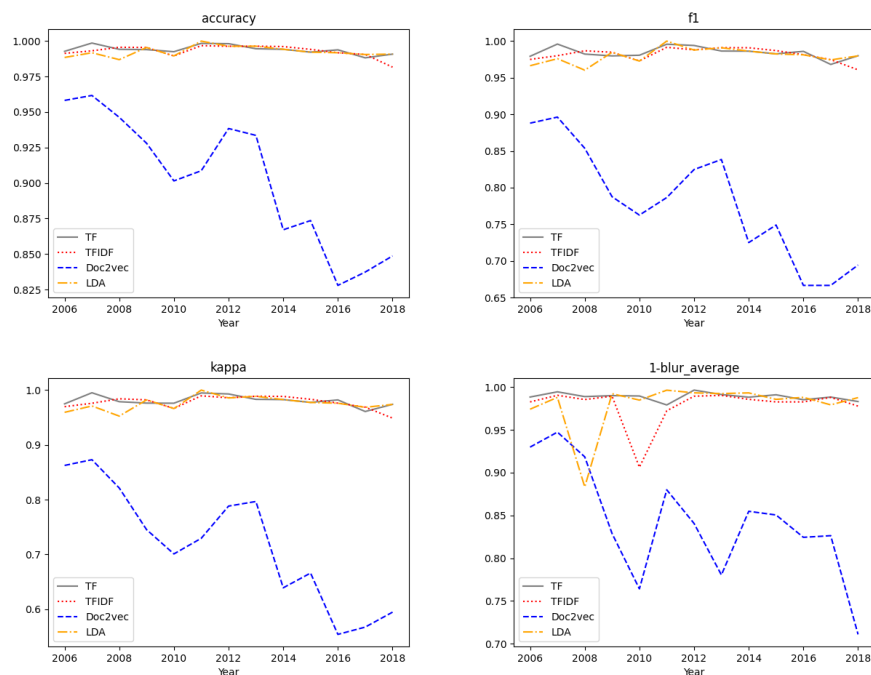


Fig. 2.16 Evaluation of the four vectorizations using accuracy rate (upper left), F1 score (upper right), Kappa score (lower left), and average of 1 - *Blurring* (lower right).

2.6 Discussion and summary

In this chapter, I introduce convergence measures based on text data at different levels of aggregation, i.e. micro-micro, micro-macro and macro-macro levels, both across- and within-classes. I make methodological contributions to addressing the practical problems of measuring individual specific convergence measures against a class, and the convergence measures across classes and within a class.

I apply the convergence measures to real-world data, i.e. 10-K reports of insurers and banks. I find potentially promising results in the application to systemic risk, e.g., micro-macro level convergence can be used to predict systemic risk in the banking industry: if a bank's risk disclosure was converged to the insurance industry in 2018, it would experience high across-industry systemic risk during the Covid-19 Pandemic; if a bank's risk disclosure was similar to a prototypical banking industry document in 2018, it would have lower within-industry systemic risk during the Pandemic, implying that focusing on industry-idiosyncratic issues reduces firm's systemic risk. Nonetheless, it is very notable that the results are highly dependent on the vectorization methods used.

My application to real-world data shows the importance of the choice of vectorization methods. The choice of vectorizations depends on the research questions, or is often

determined by the ability of the vectorization to correctly classify documents (Krzyszewska et al., 2022; Leelaprute and Amasaki, 2022) for labeled data, or purity of clustering for data without labels (Singh and Shashi, 2019). However, when applied to my data, such criteria do not give us enough evidence to choose a particular vectorization method, because all of them have reasonably good classification performance. Hence the need arises for other (potentially qualitative) criteria for evaluating the appropriateness of a document vectorization method for addressing a specific research question, which is a promising future work direction.

Chapter 3

Digitalization, changes in management attention and convergence across industries

3.1 Introduction

In this chapter, I study the industry convergence through the lens of Attention-Based View, and specifically the impact of *digitalization* on the contextual antecedents of the convergence in management attention.

The links between management attention and behaviors are well-documented (Cohen et al., 1972; Cyert and March, 1963; Joseph and Wilson, 2018; March et al., 1976; March, 1958; Ocasio, 1997). A rich literature has flourished to unfold the specifics of such links (e.g., Eggers and Kaplan, 2009; Joseph and Gaba, 2020; Tripsas and Gavetti, 2000; Vuori and Huy, 2016) with a key tenet that the antecedents of management attention should be sought in the context of decision – e.g., the material, social and cultural stimuli that impinge upon decision activity (Ocasio, 1997). Extant studies have shed light on management responses to specific, often focused in scope, changes, in the context of decision (Chen and Miller, 2007; Dutt and Joseph, 2019; Nigam and Ocasio, 2010; Ocasio and Radoynovska, 2016; Piezunka and Dahlander, 2015; Sullivan, 2010). Yet, we know little about how management attention changes in response to pervasive change. Digitization and connectivity – hereafter, digitalization – provide an example of pervasive change and form the context of this research. Digitalization is widely considered as a pervasive transformational force for people, firms, societies and institutions. It is often identified as a key strategic priority for firms and even as a driver of the “fourth industrial revolution” (McKinsey 2020, World Economic Forum 2016).

Hence, my research question is: do managers across firms and across industries respond in the same way, or differently, to digitalization? And if so, how?

A rich “digitalization literature” has developed in organization and management studies, which offers insights on the implications of digitalization on organizations and their environment (e.g., Bailey et al., 2019; Brynjolfsson and McAfee, 2014). This literature allows us to identify in a systematic way the effect of digitalization on the contextual antecedents of management attention. Building on these insights, my key claim is that digitalization leads to convergence in management attention across firms and industries. I say that there is convergence in management attention, when managers direct their attention to opportunities and concerns that are not firm-specific but common across firms and industries. I then argue that digitalization leads management to focus on their business models and that this focus is in itself a further source of convergence in management attention. I test my hypotheses in the context of banks, insurers and pharmaceutical companies. All these industries have been impacted by digitalization (e.g., Eling and Lehmann, 2018; Sehlstedt et al., 2016).

My dataset contains 10-K submissions to the SEC (Securities and Exchange Commission) of all the 934 banks, 221 insurers and 223 pharmaceuticals listed on the New York Stock Exchange from 2006 to 2018. To test my hypotheses, I focus on the two most substantial sections of those reports: the Business and Risk Factors sections. The Business section includes a company’s main products and services, what markets it operates in, recent events, competition the company faces, regulations that apply to it, labor issues, special operating costs, or seasonal factors (SEC, 2021). The Risk Factors section lists the most significant risks that apply to the company or to its assets. Using 10-Ks as a source of systematic information is increasingly common in management research (e.g., Dutt and Joseph, 2019; Nadkarni and Chen, 2014). Leveraging these data, I measure convergence in management attention by a novel measure which I call “*Blurring*”. Specifically, *Blurring* quantifies the lack of confidence of a trained text classification algorithm when faced with the task of classifying a firm’s 10-K document into an industry represented in my dataset – i.e., banks, insurance, or pharmaceuticals. Digitalization is measured using word frequency in 10-K reports of an algorithm-expanded list of keywords on digital technology, which was developed and validated with a panel of experts. Similarly, managers’ focus on business models is measured by the word frequency in 10-K reports of the keywords that reflect a focus on business model. Such keywords are derived from the rich business model literature (e.g., Massa et al., 2017). The hypotheses are tested, separately, on text data from the Business and the Risk Factors sections of 10-K reports, which can be seen as broadly reflecting the concerns of different organizational functions. Model estimates are consistent with my hypotheses, for both the Business and the Risk Factors sections.

My contribution to the literature is fourfold. First, I provide theoretical and empirical evidence that digitalization transforms management attention in a systematic way. I show that management attention becomes less idiosyncratic and firm-specific, to embrace more common concerns and priorities that span firm and industry boundaries. I also show that a significant part of this convergence in attention is due to a focus on business models and on their rewiring within digital ecosystems. Jointly these results imply that digitalization changes the content of attention. Management attention has different dimensions, including vigilance, quality, quantity etc. (Ocasio, 2011; Weick and Sutcliffe, 2006), which are already studied in literature. In this chapter, I complement extant literature by highlighting a less-studied dimension - context of decision. Management attention is contextually situated, and digitalization impact management attention by changing the context of decision when addressing management responses.

Second, my study shows that the impact of digitalization on the convergence of management attention operates across industry boundaries. By revealing the systematic implications of pervasive change, this result complements existing studies which have focused on the relationship between specific changes in the environment and management attention in individual firms or industries (e.g., Dutt and Joseph, 2019; Eggers and Kaplan, 2009; Hoffman and Ocasio, 2001; Maula et al., 2013).

Third, consistently with the above, I find that convergence engendered by digitalization also takes place across functions within organizations. Specifically, I show empirically that digitalization impacts in the same way the convergence of management attention across industries as manifested in the management agendas of business functions and risk management functions. Thus, I nuance extant insights which have found that organizational functions respond differently to the same stimulus (Joseph and Gaba, 2020; Ocasio and Joseph, 2005; Rerup, 2009) by demonstrating the systematic effect within organizations triggered by digitalization.

Fourth, I make a methodological contribution to the management literature on text analysis by suggesting a novel method for measuring similarity across sets of documents, via the new metric *Blurring*, which captures the difficulty in classifying a document into distinct categories (e.g., industries). In my context, *Blurring* measures the extent to which crossing industry boundaries is associated with step changes in the content of management attention. More broadly, this metric can be used in management applications where a measure of text similarity is needed. The idea behind *Blurring* is quite general; while in my paper I calculate *Blurring* by applying logistic regression on word frequencies, one can equally well apply a different classifier (e.g., a neural network) on a more advanced vector representation of documents (e.g., via word embeddings).

Finally, I draw implications for management and policy. While I argue that convergence in management attention might make communication and cooperation easier across firms and industries, I also recognize that it might increase “groupthink” and, ultimately, systemic risk. Insofar, I highlight some of the strategic and organizational trade-offs between convergence and differentiation, which policy makers and managers must grapple with.

3.2 Literature review and hypotheses development

3.2.1 Contextual antecedents of management attention

The fact that management attention is contextually situated and socially structured is an established tenet of the management cognition literature. For example, Nigam and Ocasio (2010) reveal how attention to Clinton’s healthcare reforms created opportunities for cognitive realignment and transformation in institutional logics, while Dutt and Joseph (2019) show the effects of regulatory uncertainty in the renewable electricity industry on corporate’s strategic agendas. Piezunka and Dahlander (2015) show that “crowding” derived from distant search narrows the attention of organizations. Li et al. (2013) find that search selection and intensity both independently and jointly influence top management attention to innovation. Chen and Miller (2007) investigate organizational responses to the threat of bankruptcy and posit that search behaviors depend on whether managers focus their attention on bankruptcy, aspirations, or slack, and that the focus of attention depends further on organizational situations. Sullivan (2010), by examining the formation of airline safety rules by the Federal Aviation Administration show that in the rule finalization stage, attention is guided by “urgency”, which interacts with certain institutional factors and with a priori “priority” given to different types of rules. Overall, these papers have nuanced and enriched my understanding of management attention in response to focused, or limited, changes in the context of decisions.

Studies that have explicitly linked technological change and changes in management attention are rare. This literature has mainly focused on information technology (Bailey et al., 2019; Christen et al., 2009; Heavey et al., 2020) and on changes at the level of a individual firm (e.g., Dutt and Joseph, 2019; Eggers and Kaplan, 2009; Maula et al., 2013) or a single industry (e.g., Hoffman and Ocasio, 2001). Christen et al. (2009) show that advances in information technology can account for the pressure on management to become more focused. Likewise, Mabey and Zhao (2017); Newell et al. (2001) have argued that digital technology might enable organizations to become more specialized. On the other hand, other researchers have argued that digital technology might broaden the scope of the

firm because of the increased potential for coordination and cooperation across multiple “units” (e.g., Leonardi and Vaast, 2017; Yoo et al., 2012) and across digital ecosystems (e.g., Adner and Kapoor, 2010; Jacobides et al., 2018). In short, the implications of technological change on management attention remain a theoretical puzzle.

3.2.2 Hypothesis development: the impact of digitalization on management attention

The technical process of digitizing and connecting knowledge and products within and across firms impacts the organizational context in several ways. “Digitalization” literature has revealed some of the transformations that digitalization brings to organizations and their environments (e.g., Bailey et al., 2019; Barrett et al., 2012; Baum and Haveman, 2020; Brynjolfsson and McAfee, 2014; Cohen and Tripsas, 2018; Faraj et al., 2011; Kellogg et al., 2020; Lanzolla et al., 2021). Digitalization takes various forms, including changing operation processes, integrating organizations into ecosystems, and reshaping business models. For example, by adopting new skills in automation, artificial intelligence (AI) and Internet of Things (IoT), insurers enhance predictive analytics involving claims probability, establish customer risk profiles and underscore underwriting and pricing models (PwC, 2020b). Examples in the banking industry include round-the-clock access to online banking, cashless transactions (HCLTech, 2020). After realizing the disruptive power and enormous opportunities of digitalization, both insurers and banks reevaluate their existing business models and develop strategies to embrace new digital technology and join new digital ecosystems (Lee and Shin, 2018). For example, thanks to digital technology, the operation process of both insurers and banks is reformed to be automated, largely changing the structure of management cost; The reporting of both industries become real-time, and periodic reporting is less useful in driving investment and marketing decisions; Self-service, enabled by smart agents in banks and insurers, raises similar new risks for both industries (Deloitte, 2018).

In what follows, I draw on this digitalization literature to identify the changes in the organizational context, which are engendered by digitalization, and to elaborate on how such changes, cumulatively, impact upon the antecedents of management attention.

First, the adoption of digital technologies is often accompanied by the onboarding of “digital experts” (e.g., McKinsey, 2020) – e.g., software engineers, data scientists, machine learning experts, etc. It is not uncommon to read that more and more companies from all traditional sectors are becoming “technology companies” because of the onboarding of digital skills and practices related to the adopted digital technologies (Fisher, 2018). As firms

increase the number of their “digital experts”, the organizational context changes. According to Brown and Duguid (1991, 2001), networks of (e.g., digital) practice cut horizontally across vertically integrated organizations and extend far beyond the boundaries of the latter. Along communities of practice, practices are shared, and information and knowledge can flow easily across organizational contexts (Brown and Duguid, 1991). Furthermore, digital skills tend to be less firm-specific than the legacy skills they replace (Lanzolla et al., 2021). It follows that the onboarding of digital experts related to digitization and connectivity is reflected in the dissemination of digital knowledge that cuts across organizational contexts.

Second, the digitalization of a firm’s knowledge base increases knowledge modularization and standardization (Baldwin and Clark, 2003; Nambisan et al., 2017; Pentland and Feldman, 2007; Schilling, 2000). Although pockets of domain-specific knowledge remain and perhaps become even “stickier” (e.g., Lanzolla et al., 2021), as firms leverage more standardized knowledge to design, produce, and support products and services throughout the organization and its value chain, they share more data and processes across organizational boundaries (Yoo et al., 2012). Digitized knowledge has new affordances such as transferability and speed. Transferability means that digital technologies allow their users to readily convey changes in one part of the system to other parts of the system (e.g., Boland et al., 2007; Forman and van Zeebroeck, 2019; Leonardi and Bailey, 2008). Speed means that digital technologies allow for the instant transmission of real-time data across a wide network of generative and connected “products” by increasing responsiveness and by reducing latency (e.g., Lyytinen et al., 2016; Teece, 2018). It follows that knowledge modularization enabled by digitization and connectivity is reflected in faster and more pervasive sharing of knowledge across organizational contexts.

Third, digitalization brings in new logics idiosyncratic to the use of digital technology. Along with onboarding digital skills, digitalization does not leave organizational contexts ‘as they are’ (e.g., Lee and Berente, 2012). Leonardi and Barley (2010); Nigam and Ocasio (2010); Orlikowski and Scott (2008) show that digital technologies create their own new realities or logics across firms. For example, Barrett et al. (2012) show that the introduction of drug-dispensing robots in two UK hospitals reconfigures the boundary relationships among pharmacists, technicians and assistants, and the highly mobile nature of digital materiality allows the two hospitals to share similar tuning processes and outcomes. Jarzabkowski et al. (2015) study how financial objects are constituted in global reinsurance markets and demonstrate how digital technologies enable the introduction of new products that bundle together risks from different territories, industries, and organizations. Such products become de-contextualised, with their evaluation primarily dependent on technical knowledge. It

follows that digitization and connectivity bring new logics that cut across organizational contexts.

Thus far, I have shown that the digitalization is likely to affect organizational context in at least three ways – pervasive digital skills and practices; sharing of non-firm specific modularized knowledge; new logics related to the use of digital technology – and that these changes, cumulatively, are in the direction of creating more commonalities across organizational boundaries. Hence, I argue that, irrespective of a firm's industry, with a higher level of digitalization in the organization, managers are increasingly likely to direct their attention towards more common rather than firm-specific issues, a phenomenon that I call convergence in management attention.

H1: Digitalization is positively associated to convergence in management attention across firms in different industries.

Digitalization extends beyond the technical processes of adopting digital technologies and has been linked to profound transformation in business contexts (e.g., Porter and Heppelmann, 2015) as well as economies, institutions and societies (Forum, 2016). For instance, in PwC's 21st CEO Survey, 81% of the respondents agree that technological progress, especially digitalization, will fundamentally change their organization (PwC, 2020a). Similarly, KPMG (2020) shows that 67% of organizations are accelerating their digital transformation strategies. A vast literature has emerged to document and investigate these changes. For instance, Zott and Amit (2007) show that e-businesses require new value creation and value capture mechanisms. More broadly, Teece (2010) and Amit and Zott (2012) show that digitalization opens up many more and new possibilities for performing companies' activities and for rewiring a firm business model, including by leveraging digital ecosystems (Adner and Kapoor, 2010; Jacobides et al., 2018). Lanzolla and Markides (2021) argue that in the new digital context a focus on internal or external interdependencies among activities may lead to competitive advantage even in conditions of resource homogeneity and low entry barriers.

Following Christensen, Raynor and McDonald's seminal work (Christensen et al., 2016), the strategy and entrepreneurship fields have equated the digital context to a catalyst of (entrepreneurial) innovation (e.g., Masters and Thiel, 2014; Teece, 2010): new business models might be based on exploiting digital distribution channels, creating and serving new customer demand, establishing new forms of customer engagement and relationships, or any combination of the three. Concepts such as "disruption" have become popular not only for the corporate agenda, but also for the broader public.

Overall, the opportunities and challenges driven by digitalization put pressure on managers in established companies to adapt or transform their business models, or even design

completely new ones. In this context, one can assume that digitalization has a *de facto* influence on every industry as it impacts corporate strategies and challenges existing business models to be reconsidered and adapted. As such, business models may feature high in management attention and priorities across firms and industries.

H2: Digitalization is positively associated to management attention on business models.

Both phenomenological evidence and academic literature show that the rewiring of business models is often associated to the emergence of digital ecosystems. Rewiring business models by leveraging digital ecosystems requires firms' joint value co-delivery (Grover and Kohli, 2012) and this in turn implies sharing of resources (Faraj et al., 2011; Lusch and Nambisan, 2015), and knowledge (Grover and Kohli, 2012). For instance, Boudreau (2010) and Parker et al. (2016) show that firms involved in ecosystems exhibit higher levels of collaboration. Yoo et al. (2012) find that firms that join ecosystems find themselves involved in more common problem solving. Jacobides et al. (2018) show how dependencies in ecosystems are based on various types of complementarities —super-modular or unique, unidirectional or bidirectional — which determine the ecosystem's added value.

In the context of financial industries, the FinTech ecosystem is an example of joint value co-delivery. In the FinTech ecosystem, financial institutes in different industries, together with all the other stakeholders, contribute to the innovation, stimulate economy, facilitate collaboration and competition, and ultimately benefit consumers (Lee and Shin, 2018; Maiti and Ghosh, 2021). Financial institutes are not able to provide convenient services for customers to choose without (a) technology developers' provision of digital platforms and favorable technical environment; (b) governments' provision of a favorable regulatory environment; (c) open customer-driven innovation.

It follows from the arguments above that participation in ecosystems enables commonalities and shared sense-making processes. Jointly, these arguments imply that management attention on business models and ecosystems leads to embracing common value co-delivery, common sense making, etc. and this is likely to lead to convergence in management attention across firms in different industries.

H3: Management attention on business models engendered by digitalization is positively associated to convergence in management attention across firms in different industries.

Thus far, I have theorized that digitalization impacts convergence in management attention through two distinct paths. First, via the technical implications – e.g., adoption of common new practices and skills – of the processes of digitizing and connecting existing organizations (H1). Second, via the broader strategic implications of such pervasive digitization and connectivity processes, which often require a rewiring of firms' business models in the

emergent digital ecosystem (H2 and H3). Taken together these hypotheses imply that digitalization leads to convergence in management attention because it induces concurrent strategic and technical changes. Figure 3.1 summarizes my hypotheses.

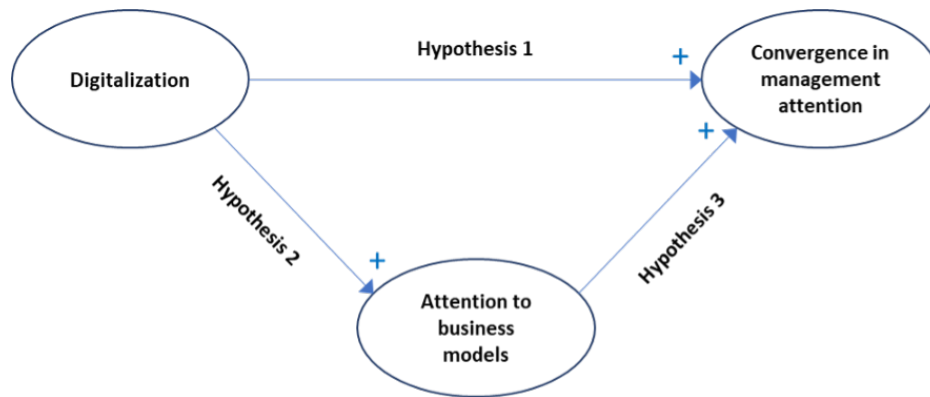


Fig. 3.1 The impact of digitalization on management attention.

3.3 Data and methods

3.3.1 Data sources

The text data used to construct the metrics come from firms U.S. 10-K reports, downloaded from Filings Expert. The 10-K Form is an annual report required by the U.S. Securities and Exchange Commission (SEC), giving a comprehensive summary of a company's financial performance, including descriptions of the firms business and the risks it faces. In the management literature, 10-K submissions have been viewed as providing a reasonable reflection of managers attention in different areas and functions of the organization (e.g., Dutt and Joseph, 2019; Flammer and Bansal, 2017; Guo et al., 2017; McKenny et al., 2018; Philippe and Durand, 2011).

My data cover 13 years (2006 to 2018, year of issue), including a total of 8331 observations (number of firm/years) from three industries. The data include 934 banks, 221 insurers, and 223 pharmaceutical companies. These comprise eleven sub-industries, as shown in Table 3.1. Initially, companies are categorized according to their NAICS codes from the Orbis database. Subsequently, those firms categorized as offices of bank holding companies, are reassigned to either the existing categories of credit banks or investment banks; or to a new category that I call mixed banks. This classification is carried out depending on the business that they primarily run, using data from Bloomberg's company profile (Bloomberg, 2021). Specifically, firms are categorized as mixed banks if they run the business of both Credit

Banks (e.g. making loans, offering checking and savings accounts and certificates of deposit, etc.) and Investment Banks (e.g. underwriting new debt and equity securities, helping with selling securities, driving mergers and acquisitions, etc.) according to their company profile.

Table 3.1 Industries and number of observations (firm/year 10-K reports)

Insurers		Banks		Pharmaceutical Companies	
Life Insurers	373	Credit Banks	4722	Pharmaceutical Preparation Manufacturing	700
Property and Casualty (PC) Insurers	746	Investment Banks	50	In-Vitro Diagnostic Substance Manufacturing	16
Full Line Insurers (running both Life and PC insurance business)	153	Mixed Banks (holding companies having both Credit and Investment operating banks)	1320	Medicinal and Botanical Manufacturing	48
Reinsurers	124			Biological Product (except Diagnostic) Manufacturing	71

SEC rules require that 10-K reports follow a set order of topics. For instance, since 2005 the SEC requires information about risk management in the *Risk Factors* section, and about strategy and business model in the *Business* section of the report (SEC and Commission, 2011). In my analysis, I only consider the *Risk Factors* and *Business* sections of the 10-K reports, as they are the sections that provide most of the discursive content of the reports. Furthermore, I construct metrics and carry out analyses separately for the *Risk Factors* and for the *Business* sections. I consider that these two sections broadly reflect the concerns of the risk management and the strategy functions, respectively.

All the text data are pre-processed by the following standard steps (see e.g., Anandarajan et al., 2019): (a) Stop words (e.g. “and”, “to” etc.) are removed because they do not carry useful information for my analysis and appear with high frequency. (b) Bigrams are constructed, such as “real estate”; analysis treats bigrams as one meaningful unit rather than two unrelated words. (c) All the words and bigrams are lemmatized. Lemmatization refers to removing “inflectional endings only and to return[ing] the base or dictionary form of a word” (Stanford, 2009a).

We use term frequency vectorization method. As discussed in Chapter 2, when using TFIDF, more weight is put on rarely used terms. TFIDF is not appropriate to reflect management attention, which may bias the informative features of the managers’ attention by focusing on rare terms.

The data sources for firm-specific control variables are from Compustat.

3.3.2 Dependent variable

Convergence of Attention. I measure the convergence of attention across firms in distinct industries by a construct I call *Blurring*, a new text similarity measure of one document against a class of documents. *Blurring* for a specific company is understood as *the difficulty in identifying the industry or subindustry it belongs to, using a classification algorithm applied on the text data*. For example, suppose a classification algorithm has been trained to decide whether a company belong to Class A or B, based on its 10-K report in a given year. Assume that, in 2006, the classification algorithm assigns a probability of 99% to the event that a particular company belong to Class A is indeed identified as A (the ‘predicted probability of the true type’). Furthermore, assume that, in 2018, the predicted probability of the true type, for the same company, has dropped to 51%. Even though in both years the company is correctly classified as A, the user of the classification algorithm would be less confident of this conclusion in 2018 compared with 2006. In that case, I say that the company is more blurred in 2018 compared with 2006. (I would make the same statement if in 2018 the predicted probability of true type had decreased to 40%, leading to mistaken classification of the company as Class B.) Hence, I define the *Blurring* of a firm, in a given year, as one minus the predicted probability of true type. This means that in my example, the *Blurring* of the company in 2006 was 1%, while in 2018 it was 49%.

In this chapter, I use logistic regression as a classifier, applied on normalized document word frequencies. Logistic regression is chosen here due to its simplicity and common use. I note that the definition of *Blurring* can be applied in exactly the same way for any classification method that produces predicted probabilities, e.g. a neural network with a softmax output layer (Géron, 2019). A more formal definition of *Blurring* and details on the algorithm by which it is constructed can be found in Appendix A.

The process of calculating *Blurring* takes place in two distinct scenarios, each of which corresponds to a different level of aggregation. *Blurring* in each scenario is calculated separately twice: first considering only the Risk Factors section of 10-K reports and, second, the Business section of the reports (hence two classification analyses with different underlying data are carried out).

Scenario 1: Convergence in attention across different industries. There are three categories, representing insurance, banking and pharmaceuticals. The *Blurring* corresponding to firm i and year t is denoted by $BlurRiskAcross_{it}$ (*Risk Factors* Section) or $BlurBusAcross_{it}$ (*Business* Section).

Scenario 2: Convergence in attention within the same industries. For Insurers, Banks and Pharmaceutical companies, there are four, three and four categories respectively, with

each category corresponds to a sub-industry, as is shown in Table 3.1. A company's *Blurring* is calculated within each industry that it belongs to. The *Blurring* corresponding to firm i and year t is denoted by $BlurRiskWithin_{it}$ (calculated from the Risk Factors Section) or $BlurBusWithin_{it}$ (calculated from the Business Section).

For each of the above scenarios, the accuracy rates of the classification algorithm used to calculate *Blurring* are presented in Table 3.2. These rates refer to the percentage of accurately classified observations (company/years). Their high values ($> 90\%$) of *Blurring* Across the three industries indicate good performance of the logistic regression model. The values of *Blurring* Within each industry are much lower, because sub-industries within the same industry are more similar compared to different industries.

Table 3.2 Accuracy rates of classification

	Across Insurers, Banks and Pharma	Within each industry
Risk	98.59%	52.30%
Business	98.93%	49.83%

Figure 3.2 shows the time trends of average *Blurring* in Scenario 1 (*Blurring* across different industries), together with the 95% confidence intervals. In each plot the horizontal axis shows the year, while the vertical axis represents the *Blurring*. I report the *Blurring* by the three industries in different columns, and by the two sections, *Risk Factors* and *Business*, in different rows.

The increasing trend of *Blurring* shows that, over the years, the classification algorithm places a decreasing level of confidence on a company being classified as the true type.

3.3.3 Independent variables

To measure the attention to digitalization and business models, I first construct the keyword lists on them in the same way as follows:

1. I construct *seed keyword* lists on business model and digitalization.

The seed list of keywords on digitalization was created as follows. My construct measurement begins with the definition of the concept of digitalization. Digitalization was defined as an unfolding technology-related phenomenon that is triggered by specific enabling digital technologies and poses several novel opportunities and challenges at multiple levels including for people, firms, societies and institutions. These enabling technologies provide the constructs dimensionality. Therefore, to measure management attention to digitalization, I developed an exhaustive list of enabling

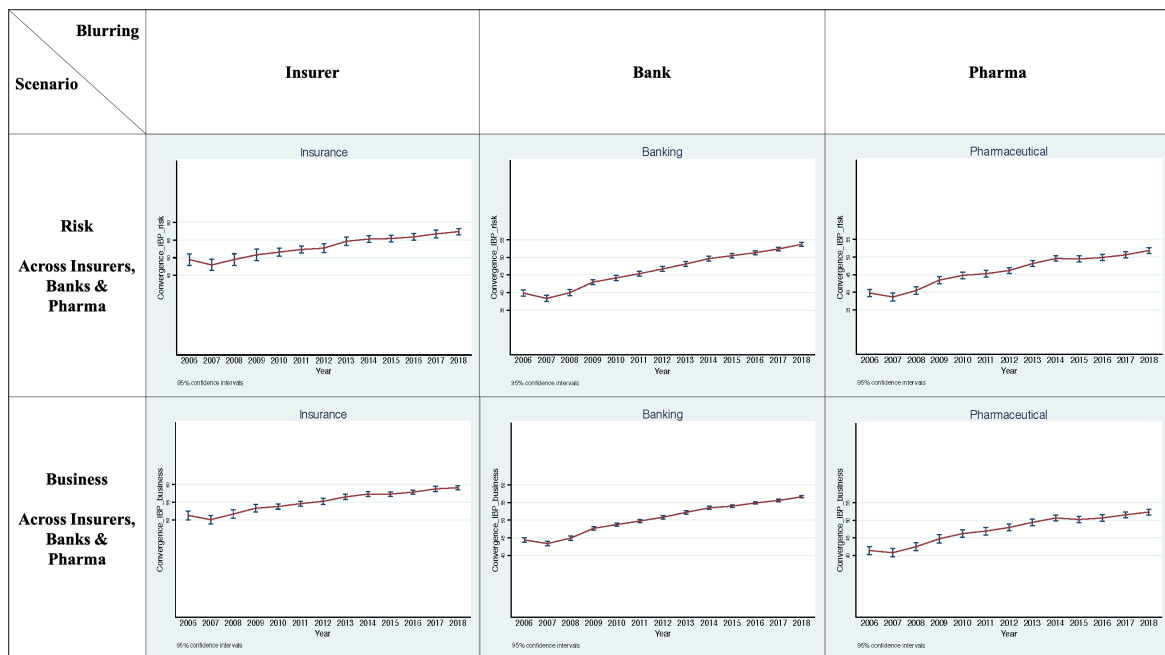


Fig. 3.2 *Blurring* across insurers, banks and pharmaceutical companies.

The red lines show the average of *Blurring*, and the blue interval shows the 95% confidence interval. I report the *Blurring* in the Scenario 1 (across different industries) by the three industries (Insurers, Banks and Pharma, row), and by the two sections (*Risk Factors* and *Business*, column). All the sub-figures are of the same vertical scale, ranging from 35% to 60%. The horizontal axis shows the years, ranging from 2006 to 2018.

digital technologies and related keywords to capture digital orientation construct. To further validate this work list, I created an online questionnaire and asked CEOs and digital transformation experts to assess these words and suggest alternative words or additional words to capture these technologies. The panel of CEOs and experts included CEOs from firms in the IT industry as well as partners in IT consulting firms. Their suggestions mainly corroborate my list of words without major changes. After the validation of the keywords, I created concordance lists, meaning list of narratives that included the focal keywords and I examined the construct. Finally, I used them as the seed keywords on management attention to digitalization. 9 keywords are included in the seed keyword list on digitalization.

Regarding the seed list of keywords on business model, I manually identify the keywords based on the seminal review paper on business model by Massa et al. (2017). I read all the definitions of business model in the paper and include 90 keywords in the seed keyword list on business model.

All the seed keywords can be found in Table 3.3.

2. I *extend* the keyword list based on the seed keyword list. To do this, I find all words in vocabulary of my data, which are semantically close to the seed words. This is achieved by the following process.
 - (a) I represent each single word in the vocabulary, as a vector, using Word2vec embeddings. Word2vec embeddings of words enables us to assess the semantic distance between two words (Mikolov et al., 2013a,b).
 - (b) For each word, I then calculate the cosine similarity of its embedding to each of the embeddings of the seed keywords. If the highest such similarity is more than 0.5, I include that word in the extended list of keywords. This leaves us 91 keywords on digitalization, and 707 keywords on business models.
 - (c) I *manually screen* each word in the extended list. I remove the words that are too generic or obviously unrelated. The final keyword list on digitalization includes 61 keywords, and on business model 276. The full list of keywords is given in Appendix B

My measure of digitalization and business model reflects the attention of the organization to the topics. Hence, the measure provides a proxy for the extent that individual firms in different industries and years are impacted by attention to digitalization or business model, given that the literature has shown a robust link between attention and resource allocation (Vanacker et al., 2020). Measuring directly the full breadth and depth of digitalization or

Table 3.3 List of seed keywords on digitalization and business model

Seed keywords on digitalization
analytic, AI (artificial intelligence), automation, internet, cyber, social media, digit, digital, big data
Seed keywords on business model
business unit, function, value creation, market, cooperation, environment, connection, activity, economic value, makes money, value chain, content, structure, governance, opportunities, relationship, operation, interdependency, customer, suppliers, partner, stakeholder, department, staff, value, money, economic logic, cost, value capture, value creating, value creation, interdependency, feedback loop, monetizing, system, activity system, value proposition, product, supply chain, financial model, benefit, distribution, new venture, venture, profitable, strategy, components, profit, process, performance, resources, efficiency, effectiveness, profitability, offer, segments, market segment, customer segment, architecture, revenue, marketing, deliver, relationship capital, internal value, organization, design, value capture, value proposition, value delivery, natural resource, social resource, economic resource, ecosystem, logic, value network, product offering, service offering, articulation, internal, external, exchange, regulation, incentive, price, policy, socioeconomic, output system, capture, proposition, constellation

attention to business model, across time, for such a heterogeneous and large database of companies is practically not achievable. I believe that my approach is a reasonable one to address this measurement problem.

Finally, while my measures for convergence – i.e., *Blurring* – and attention to digitalization and business model – are constructed starting from the same dataset, this is not a source of endogeneity. First, to avoid endogeneity, prior to calculating the firm/year-specific values of *Blurring* for regression analysis, I remove all keywords used to construct independent variables; in any case, such words only form a tiny fraction of the vocabulary. Second, the two metrics capture very different features of the data, that is, digitalization and business model reflect the frequency of words relating to digitalization and business model in individual documents, while *Blurring* relates to the similarity between the word frequencies of individual firms' reports v.s. sets of reports from a different industry.

Attention to digitalization. The attention to digitalization in each company is defined via the word frequency of all the word in the list of keywords on digitalization and the document of each company per year.

Assume that there are n words in total in the vocabulary and, for simplicity, assume that the first $k < n$ of those are keywords. Then, for a document with word frequency vector $D_{it} = (D_{it,1}, \dots, D_{it,n})$, the word frequency equals

$$Digit_{it} = \frac{\sum_{j=1}^k D_{it,j}}{\sum_{j=1}^n D_{it,j}} \quad (3.1)$$

that is, it is the proportion of the sum of the *frequency of the keywords* out of the *frequency of all the words* in the document.

When calculating the attention to digitalization for a company in a given year (denoted as $Digit_{it}$), I consider the frequency between the list of keywords on digitalization and a document containing both Item 1A *Risk Factors* and Item 1 *Business* from the respective 10-K form for this firm/year. I show digitalization trends in Figure 3.3a by plotting the average level of $Digit_{it}$ in each year by industry. There seems to be a slight decrease in early years, with 2012 appearing to be an inflection point, after which there is a clear increasing trend in both insurance and banking industries, while a relatively flat trend in pharmaceuticals.

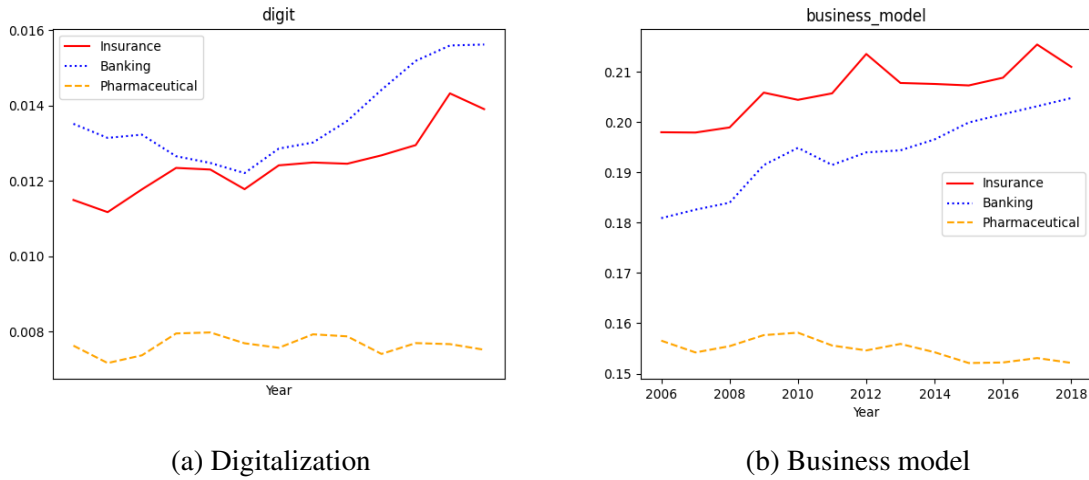


Fig. 3.3 Average level of attention to digitalization and business model in each year across companies.

Measured by the word frequency in 10-K reports (both *Business* and *Risk Factors* Sections).

Attention to business models and ecosystems. I measure management attention to business models and ecosystems using the same method employed to measure digitalization. Specifically, attention to business models and ecosystems is defined via word frequency between the list of relevant keywords on business model and the document of each firm per year, including both Item 1A *Risk Factors* and Item 1 *Business* from the respective 10-K form. I denote the resulting variable by $BusModel_{it}$.

As mentioned above, the list of seed keywords on business models and ecosystems is given in Table 3.3 and was derived by consulting the business model literature (e.g., Massa et al., 2017). Similar to the construction of $Digit_{it}$, in the calculation of the dependent variable *Blurring*, words relating to business models and ecosystems were removed to avoid endogeneity.

3.3.4 Control variables

Control variables are firm/year-specific, and pertain to firm size, financial performance, and the sub-industries that companies belong to. I also control year fixed effect.

The first control variable is size. I use the amount of total book assets to represent the size of companies (Dang et al., 2018), coming from the Compustat database (Compustat, 2021). This value is extremely skewed, so it is log-transformed. The covariate $\ln asset_{it}$ is included in the regression analysis as a control, because size is associated with diversification. Fama and Jensen (1983) argue that large companies are more diversified than small companies. In addition, large companies form economies of scale, and have more resources (e.g. negotiating power) to diversify into other industries.

Performance is controlled because it is associated with diversification, which can increase the similarity between different industries. I use Return on Equity as the proxy of performance for insurers and banks. Data on Return on Equity come from Compustat. I define ROE_{it} as the Return on Equity of the previous period (i.e. $t - 1$), to avoid endogeneity.

The sub-industry the firm belongs to is controlled using dummies when we use regression models other than fixed-effect panel data regression later. The sub-industry dummies are denoted as $life_i$ (equal to 1 if the company is a life insurer, and 0 otherwise), pc_i (property and casualty insurer), $fullline_i$ (full line insurer), $reinsurance_i$ (reinsurer), $credit_i$ (credit bank), $investment_i$ (investment bank), $mixedbank_i$ (mixed bank), $vitro_i$ (in-vitro diagnostic substance manufacturing firm), $medicinal_i$ (medicinal and botanical manufacturing firm), and $biological_i$ (biological product manufacturing firm). Pharmaceutical preparation manufacturing companies are the baseline.

The use of similar non-informative language in the 10-K reports, which is heavy with cliches, may also leads to convergence. Such non-informative is measured using boilerplate, defined as the standardized expressions carrying little meaning (Lang and Stice-Lawrence, 2015). The resulting variables, $BoilerRisk_{it}$ (percentage of boilerplate in *Risk Factors* section) and $BoilerBus_{it}$ (percentage of boilerplate in *Business* section), are calculated by the following process based on Lang and Stice-Lawrence (2015).

- I count all tetragrams contained in each document in my sample, where a tetragram is an ordered group of four words within a single sentence. I aggregate these counts by year of issue.
- I identify tetragrams that occur in at least 30% of the documents or on average at least 5 times per document across all the three industries in a year. The identified tetragrams are considered to reflect non-informative ('boilerplate') language.

- Then the percentage of the common boilerplate tetragrams out of all the tetragrams in each document is calculated, and set equal to *BoilerRisk_{it}* or *BoilerBus_{it}*.

Descriptive statistics (before standardization) and definitions of variables can be found in Table 3.4, and correlation matrix can be found in Table 3.5.

Table 3.4 Descriptive statistics (before standardization)

Variable description	Variable	N	Mean	Median	Standard deviation
<i>Blurring of Risk Factors</i> across the three industries	<i>BlurRiskAcross_{it}</i>	8778	66.833	65.080	6.311
<i>Blurring of Business</i> across the three industries	<i>BlurBusAcross_{it}</i>	8778	46.711	47.832	9.244
<i>Blurring of Risk Factors</i> within each industry	<i>BlurRiskWithin_{it}</i>	8778	65.703	63.959	7.082
<i>Blurring of Business</i> within each industry	<i>BlurBusWithin_{it}</i>	8778	50.172	51.307	7.238
Attention to digitalization	<i>Digit_{it}</i>	8778	0.012	0.012	0.005
Attention to business models in ecosystems	<i>BusModel_{it}</i>	8778	0.189	0.191	0.030
Log transformed total book asset	<i>lnasset_{it}</i>	6069	7.529	7.560	2.393
Return on equity	<i>ROE_{it}</i>	6430	-1.076	0.072	20.242
Percentage of boilerplate in <i>Risk Factors</i> section	<i>BoilerRisk_{it}</i>	8778	0.503	0.401	0.447
Percentage of boilerplate in <i>Business</i> section	<i>BoilerBus_{it}</i>	8778	0.237	0.188	0.227

Table 3.5 Correlation matrix

	<i>BlurRiskAcross_{it}</i>	<i>BlurBusAcross_{it}</i>	<i>BlurRiskWithin_{it}</i>	<i>BlurBusWithin_{it}</i>	<i>Digit_{it}</i>	<i>BusModel_{it}</i>	<i>lnasset_{it}</i>
<i>BlurRiskAcross_{it}</i>	1.000						
<i>BlurBusAcross_{it}</i>	0.415***	1.000					
<i>BlurRiskWithin_{it}</i>	0.867***	0.225***	1.000				
<i>BlurBusWithin_{it}</i>	0.298***	0.965***	0.118***	1.000			
<i>Digit_{it}</i>	-0.313***	0.109***	-0.329***	0.183***	1.000		
<i>BusModel_{it}</i>	-0.0332***	0.447***	-0.0291***	0.495***	0.257***	1.000	
<i>lnasset_{it}</i>	0.029**	0.353***	0.0213*	0.351***	0.150***	0.387***	1.000
<i>ROE_{it}</i>	-0.071***	0.0129	-0.073***	0.036***	0.060***	0.073***	-0.0159
<i>BoilerRisk_{it}</i>	0.1284***	0.136***	0.110***	0.125***	-0.063***	-0.044***	-0.0582***
<i>BoilerBus_{it}</i>	-0.0357***	0.052***	-0.030***	0.060***	0.050***	0.043***	0.007

	<i>ROE_{it}</i>	<i>BoilerRisk_{it}</i>	<i>BoilerBus_{it}</i>
ROE	1.000		
boiler_risk	-0.050***	1.000	
boiler_bus	0.199***	-0.028**	1.000

Pearson correlation coefficients, *** p<0.01, ** p<0.05, * p<0.1.

3.3.5 Regression models

My three hypotheses, as represented in Figure 3.1, reflect my theorization of digitalization impacting convergence in management attention through two distinct paths. First, via the onboarding of digital experts, knowledge modularization, and new logics (H1); second, via the attention to and rewiring of firms' business models the emergent digital ecosystems (H2 and H3). Taken together, these arguments imply that management attention on business models

partially mediates the relationship between digitalization and convergence in management attention across firms in different industries.

To test this set of relationships, I follow the four steps below (Aguinis et al., 2017). Recall that I have four different dependent variables, corresponding to two different classification scenarios (convergence within or across industries) and two 10-K sections (Risk Factors and Business), that is, $BlurRiskAcross_{it}$, $BlurBusAcross_{it}$, $BlurRiskWithin_{it}$, $BlurBusWithin_{it}$ – denoted here collectively by $BlurXXX_{it}$.

1. I test the relationship between convergence in management attention and digitalization. For this I use panel regressions for each $BlurXXX_{it}$, with $Digit_{it}$ as the independent variable and the control variables described above. A statistically significant positive association between $BlurXXX_{it}$ with $Digit_{it}$ would be consistent with Figure 3.1 and form evidence in support of H1. The regression equation of Step 1 is shown in Equation (3.2), where α , β , γ and u are intercept, regression coefficients, unobserved time-invariant individual effect, and individual error respectively.

$$BlurXXX_{it} = \alpha_1 + \beta_{11}Digit_{it} + Control_{it}^T \cdot \beta_{12} + \gamma_{1i} + u_{1it} \quad (3.2)$$

2. I test the relationship between attention to business models and ecosystems and digitalization. For this I regress $BusModel_{it}$ on $Digit_{it}$, with the same control variables as in Step 1. A statistically significant positive association would form evidence supporting H2. The regression equation of Step 2 is shown in Equation (3.3).

$$BusModel_{it} = \alpha_2 + \beta_{21}Digit_{it} + Control_{it}^T \cdot \beta_{22} + \gamma_{2i} + u_{2it} \quad (3.3)$$

3. I test the relationship between convergence, digitalization and attention to business models and ecosystems. I regress the dependent variables $BlurXXX_{it}$ on both $Digit_{it}$ and $BusModel_{it}$ with the same control variables as before. A statistically significant positive association between the dependent variable $BlurXXX_{it}$ and $BusModel_{it}$ would form evidence supporting H3. The regression equation of Step 3 is shown in Equation (3.4).

$$BlurXXX_{it} = \alpha_3 + \beta_{31}Digit_{it} + \beta_{32}BusModel_{it} + Control_{it}^T \cdot \beta_{33} + \gamma_{3i} + u_{3it} \quad (3.4)$$

4. Finally, I need to establish whether attention to business models and ecosystems mediates the relationship between convergence and digitalization. Evidence in support of this relationship arises if the coefficient of $Digit_{it}$ in Step 3 is smaller compared to Step

1 while the coefficients of $BusModel_{it}$ in Steps 2 and 3 are significant. Furthermore, the mediation is partial, if the coefficient of $Digit_{it}$ in Step 3 remains statistically significant. In other words, the inclusion of $BusModel_{it}$ in the regression does not remove in full the ability of $Digit_{it}$ to explain variance in $BlurXXX_{it}$, such that support for H1 persists. In addition, Sobel-Goodman mediation tests (Goodman, 1960), Aroian test (Aroian, 1947), and Bootstrap analysis (Shrout and Bolger, 2002) are also employed to test whether $BusModel_{it}$ is a mediator (Preacher and Hayes, 2004).

In all the steps above, I use fixed-effects panel data regression as the baseline model, and random-effect as a robustness test. In all models, all the variables are standardized by industry, to have a mean of zero and a standard deviation of one within each industry.

3.4 Results

The regression results can be found in Table 3.6. Models (2), (4), (6) and (8) in the table report the regression results of Step 1, Model (1) of Step 2, and Models (3), (5), (7) and (9) of Step 3. In particular, Models (2), (3), (6) and (7) refer to convergence of attention within each industry; Models (4), (5), (8) and (9) to convergence of attention across the three industries. Models (2), (3), (4) and (5) relate to dependent variables calculated based on 10-K *Risk Factors* sections, while for Models (6), (7), (8) and (9) the dependent variable is calculated from the *Business* section.

Table 3.6 Regression results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$BusModel_{it}$	Risk within	Risk within	Risk across	Risk across	Business within	Business within	Business across	Business across
		$BlurRiskWithin_{it}$	$BlurRiskWithin_{it}$	$BlurRiskAcross_{it}$	$BlurRiskAcross_{it}$	$BlurBusWithin_{it}$	$BlurBusWithin_{it}$	$BlurBusAcross_{it}$	$BlurBusAcross_{it}$
$lnasset_{it}$	-0.043 (0.030)	-0.042 (0.049)	-0.033 (0.049)	0.126*** (0.034)	0.136*** (0.033)	-0.012 (0.042)	-0.010 (0.042)	0.118*** (0.032)	0.128*** (0.031)
ROE_{it}	-0.001 (0.007)	0.023** (0.011)	0.023** (0.011)	0.019*** (0.007)	0.020*** (0.007)	0.023** (0.009)	0.023** (0.009)	0.017** (0.007)	0.017** (0.007)
$BoilerRisk_{it}$		0.010 (0.009)	0.010 (0.009)	0.002 (0.007)	0.003 (0.006)				
$BoilerBus_{it}$						0.039*** (0.008)	0.038*** (0.008)	0.007 (0.006)	0.004 (0.006)
$Digit_{it}$	0.077*** (0.012)	0.064*** (0.019)	0.048** (0.019)	0.087*** (0.013)	0.069*** (0.013)	0.030* (0.016)	0.026 (0.016)	0.090*** (0.012)	0.073*** (0.012)
$BusModel_{it}$			0.207*** (0.024)		0.235*** (0.016)		0.054*** (0.021)		0.221*** (0.015)
_cons	-0.241*** (0.025)	0.199*** (0.041)	0.249*** (0.041)	-0.580*** (0.029)	-0.523*** (0.028)	0.327*** (0.035)	0.340*** (0.035)	-0.712*** (0.026)	-0.659*** (0.026)
Year FE	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled
Obs.	5458	5458	5458	5458	5458	5458	5458	5458	5458
R-squared	0.131	0.077	0.092	0.460	0.483	0.072	0.073	0.549	0.569
F	45.845	23.750	27.027	242.402	250.237	21.980	21.123	346.528	353.158
Sobel			0.0134***		0.0220***		0.0154***		0.0212***
Aroian			0.0134***		0.0220***		0.0154***		0.0212***
Bootstrap			0.0134***		0.0220***		0.0154***		0.0212***

Standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

I see that the coefficient of $Digit_{it}$ is consistently positive and statistically significant across all models (2), (4), (6) and (8). This indicates strong support for H1, indicating a robust positive association between digitalization and convergence of attention across industries.

In Model (1) the coefficient of $Digit_{it}$ is positive and statistically significant, which supports H2, as in Step 2 above.

In Models (3), (5), (7) and (9), I observe that (a) the coefficient of $BusModel_{it}$ is consistently positive and statistically significant; (b) the coefficient of $Digit_{it}$ is consistently positive and statistically significant except Model (7); and (c) the scale of the coefficient of $Digit_{it}$ reduces compared to corresponding Models (2), (4), (6) and (8), which do not include $BusModel_{it}$ as a covariate (specifically comparing (3) to (2), (5) to (4), (7) to (6), and (9) to (8)). According to the Steps 3 and 4 above, these results, taken together, provide evidence in support of H1 and H3.

Further support for the mediation model as depicted in Figure 3.1 is provided by Sobel-Goodman mediation tests, found at the end of Table 3.6, all showing that $BusModel_{it}$ is a mediator in the relationship between $Digit_{it}$ and $BlurXXX_{it}$.

I note that the results above hold consistently for both within and across industries and different organizational functions (Risk Management vs Strategy). I performed a number of robustness checks. Specifically, I re-ran the regression models (1)-(9) with the following modifications: (a) the entire sample of text documents was used, without including $lnasset_{it}$ and ROE_{it} (for which observations are not available for all company/years); (b) the U.S. internet penetration rate in each year, measured using the percentage of U.S. adults who use the internet (Pew, 2020), was included as a control; (c) robust standard errors were used when running the models; (d) models were run with only those firms that are in the NYSE for the full sampling period; (e) a higher level of regularization when calculating $Blurring$ was used; (f) logistic regression models were trained on the whole sample rather than only the documents within a single year when constructing $Blurring$; (g) random-effect panel data regressions were run for all models of Table 3.6. For all those robustness checks, the results remain consistent with all hypotheses.

3.5 Discussion and conclusions

In this chapter, I argue that digitalization systematically changes the contextual antecedents of management attention and have shown, theoretically and empirically, that these changes lead to convergence in management attention across firms and industries. Specifically, I reveal that management attention becomes less firm-idiosyncratic, embracing concerns and priorities that span firm and industry boundaries; and becomes focused on business models

and on their rewiring in digital ecosystems. My empirical findings provide strong support for these changes being systematic across firms, industries and organizational functions.

My contribution to the literature is fourfold. First, I reveal the direct link between digitalization and changes in the content of management attention, a dimension hitherto overlooked in the literature, which has instead focused on other facets of changes in attention including quality and quantity of organizational attention (Weick and Sutcliffe, 2006); bottom-up and top-down attentional processes (Laureiro-Martinez, 2014; Shepherd et al., 2017); and attentional breadth versus depth (Eklund and Mannor, 2020).

Second, extant literature has mostly addressed firm-specific or, more rarely, industry-wide changes (e.g., Dutt and Joseph, 2019; Eggers and Kaplan, 2009; Hoffman and Ocasio, 2001; Li et al., 2013; Maula et al., 2013). My findings show that changes engendered by digitalization are pervasive across industry boundaries. For instance, while digitalization-enabled convergence between banks and insurers might be hypothesized in the context of their adjacency, I find consistent results when pharmaceutical industry is also included. This points to digitalization transforming the context in which management attention is located in a profound and systematic way. Thus, I also complement the literature on industry evolution (e.g., Christensen et al., 2016; Klepper, 1997). Recent developments in this literature (e.g., Uzunca, 2018) have documented the dynamics leading to the *Blurring* and removal of intra- and inter-industry boundaries in firms' technological and customer competences. Here I show that such "*Blurring* and removal of boundaries" also manifests as convergence in management attention.

Third, my findings put forward the idea that the impact of digitalization on management attention is not only systematic across firms and industries, but it is also invariant across firms' organizational functions. This finding complements and expands on recent studies that have claimed that attention is structurally situated within organizations (e.g., Gaba and Joseph, 2013; Joseph and Gaba, 2020; Rerup, 2009); for instance, management attention may be impacted by managers' affiliation to headquarters vs. subsidiaries (Dutt and Joseph, 2019). By showing that the convergence in management attention triggered by digitalization is pervasive within organizations I suggest that the extant claim should be nuanced and understood within the contextual changes brought forward by technology diffusion.

Jointly, the findings of my research contribute to the ongoing debate on the nature of digitalization and of its implications for organizations and organizing. In this chapter, I have conceptualized digitalization as a pervasive technology with broad scope. My empirical findings show that the impact of digitalization on management attention is systematic in a way consistent with my theoretical arguments. Hence I suggest that a more complete characterization of technological change should reflect the degree of pervasiveness and the

scope of a new technology along with, e.g., industry velocity (Nadkarni and Barr, 2008), data richness (Christen et al., 2009), customer needs and demand growth (D'Aveni and MacMillan, 1990).

Fourth, I make a methodological contribution to the management literature of text analysis by suggesting a novel method for measuring similarity across sets of documents, producing a new metric that I call “*Blurring*”. The introduction of this measure was necessitated by the problem of constructing a dependent variable for my regressions, which reflects the convergence between an individual company and a whole industry. This flexible measure of convergence can be applied with any classifier that produces predicted probabilities (e.g. neural networks), and can be used to measure similarity, *Blurring*, isomorphism, etc. By constructing a rigorous measure for my dependent variable, I respond to the call for new methods to shed light on the complex relationships between context, structure and attention (e.g., Joseph and Gaba, 2020).

My findings have outright policy and management implications. On the positive side, convergence in management attention might make communication and cooperation easier within and across firms and industries. Yet, there are some pitfalls. For organizations, as bounded rationality guides resource allocation (Puranam et al., 2015; Simon, 1979), such convergence in management attention might increase groupthink and uniformity, leading to similarity in resource allocation. Concentration of resource allocation thus requires explicit new strategies for creating slack and diversity. Furthermore, policy makers should consider a potential increase in systemic risk (Chen et al., 2014; Cummins and Weiss, 2009) and develop monitoring and intervention tools which address groupthink and go beyond mere econometric and correlation measures (e.g., Brownlees and Engle, 2017).

Limitations and avenues for future research. Like all studies, my research has some limitations. One is that my data sample only cover listed companies in the U.S., as only these companies regularly disclose text reports following a consistent format. It follows that my findings should be tested also for private companies and non U.S. samples. Second, my text analysis approach has potential limitations related to whether the authors of the 10-K reports disclose exactly what they think. Third, future research might focus on the relative importance of the mechanisms proposed in this paper. Qualitative research methods might help to tackle some of these limitations.

Chapter 4

Contextual antecedents of industry convergence: text analysis of risk disclosure

4.1 Introduction

4.1.1 Problem statement

Based on the analysis in Chapter 3, two following questions are: What other factors may drive such convergence in financial disclosure? What aspect of systemic risk does such convergence reflect?

As discussed in Chapter 1, one widely accepted definition of systemic risk in the financial literature is: *the risk of disruption to the flow of financial services, (i) caused by an impairment of all or parts of the financial system; and (ii) has the potential to have serious negative consequences for the real economy (IMF et al., 2009)*. Guided by the financial and economical essence of this definition, many studies on systemic risk focus purely on quantitative economic data, e.g., stock data, to measure systemic risk (e.g., Acharya et al., 2012, 2017; Billio et al., 2012; Brownlees and Engle, 2017; Cai et al., 2018; Chen et al., 2014; Huang et al., 2009; Tarashev et al., 2016).

Nonetheless, such a narrow understanding of systemic risk, focusing purely on quantitative economic data, provides only very partial view. Such data have important limitations (Danielsson et al., 2022): historical data on extreme stresses are scarce and, where present, do not reflect general scenarios. Furthermore, it is not clear how, e.g., stock data can reflect subjective sources of system vulnerability, such as excessive trust in the market / system. Examples of excessive trust in the market / system is the widely-accepted assumptions that

the market is perfectly complete, information spread without cost, the system can adjust itself automatically to match supply and demand, and each participant's volume is trivial compared to the whole market's, while actually the real market is not so.

Furthermore, broader behavioral issues, reflected in both the financial and non-financial literatures, are potentially linked to systemic risk. For example, herding behavior has been documented, whereby individuals imitate group behavior when making economic decisions (e.g., Clement and Tse, 2005; Trueman, 1994), with such co-behaviors causing systemic risk (Haldane and May, 2011; Piškorec et al., 2014). Other examples include groupthink (Janis, 2008), management fads (Gibson and Tesone, 2001) from management literature. For example, during the 2008 financial crisis, the core activities of insurers, which attracted the main attention of regulators and scholars, were not a major source of systemic risk; instead, it was the fads for high return from non-core business, e.g., insurers' investment in mortgage-backed securities that raised the systemic risk, which were not sufficiently reflected in the traditional quantitative economic data (Cummins and Weiss, 2009).

Systemic risk has been often associated with interconnectedness (IMF et al., 2009). Learning from AIG's experience in 2008, strong interconnectedness across industry boundaries, e.g., insurers investing in mortgage-backed securities, implies that when the market experiences an extreme downside shock, the connected firms and industries (insurance and banking industries in this case) will share similar vulnerabilities and thus co-movement (Billio et al., 2012; Bushman et al., 2017). In this chapter, I will continue focusing on the convergence (one channel of interconnectedness) of insurance and banking industries as my research context. I choose this context because the convergence of these two industries has the potential to trigger turmoil like the one in 2008, and thus is worth a close look.

An interconnected world requires new ways to measure and detect risk, as we need a more timely measure to indicate systemic risk before it spreads and brings extreme consequences; or even before its formation, i.e. by detecting the antecedents of systemic risk. I use text data from financial disclosures to reflect interconnectedness and thus systemic risk, because such data reflect (convergence in) the attention or cognition of managers (Dutt and Joseph, 2019). As discussed in Chapter 1, management attention is the antecedent of risk, while other numerical measures based on financial data are mostly *ex post*. A suitable representation of management attention is formed by companies' 10-K reports, which have been widely used as a data source in management (e.g., Dutt and Joseph, 2019; Guo et al., 2017), accounting (e.g., Cazier et al., 2021), risk management (e.g., Bao and Datta, 2014), and systemic risk measurement (e.g., Bushman et al., 2017). A notable feature of my paper is my focus on convergence in attention *across industries*, which contrasts with a growing literature using financial reports to extract information about individual firms or *within* a single industry (e.g.,

Bushman et al., 2017). As systemic risk affects all parts of the real economy, measuring convergence in management attention within a single industry is not enough to reflect the potential system interconnectedness.

The research aims of this chapter are twofold. First, I develop a theoretical framework for associating systemic risk with the Attention-Based View of the firm. Through this framework, I identify constituent elements in the context of management attention, each of which may have different systemic risk implications and requires a different response. Second, I operationalize the theoretical insights by providing text-based measures of the convergence and its antecedents that I identified theoretically.

4.1.2 A Road map of this chapter

In Section 4.2, I develop a theoretical framework based on the Attention-Based View of the firm. I identify three key antecedents of convergence in attention across industries, namely: companies' description of common trends in the macro environment, managers' attention to an adjacent industry because of the threat of substitution, and management fashion of fads.

In Section 4.3, I report my text data sources and construct a text-based measure of convergence in the attention of risk managers. To build this variable, I first vectorize the risk disclosure documents using Doc2vec embeddings (Mikolov et al., 2013a,b), and then use (negative) silhouette values (Bao and Datta, 2014; Rousseeuw, 1987) to measure the convergence, by assessing text similarity between an individual firm and an industry that it does not belong to.

Measures for my proposed antecedents of convergence are constructed in Section 4.4. Specifically, as a prerequisite step, I carry out topic modelling (Blei et al., 2003) of the Business section of 10-K reports in Section 4.4.1. This enables us to identify the extent to which companies discuss topics that relate to general trends and the threat of substitution across industry boundaries. Furthermore, I use the similarity in the language of risk disclosure to pharmaceutical industry (relatively unrelated to the two industries I consider), in order to evaluate the presence of risk management fashion.

In Section 4.5, I associate convergence in management attention with its measured antecedents, via Random Forest (Breiman, 2001) regression analysis. Subsequently, we decompose the convergence of attention measure to its modelled antecedents, using Accumulated Local Effects (Apley and Zhu, 2020), feature importance (Breiman, 2001; Ishwaran, 2015), and sensitivity analysis based on Shapley values (Lundberg and Lee, 2017a,b; Tsanakas and Millossovich, 2016). I find that management fashions has the biggest impact on convergence in attention, with this result being consistent across all methods used. Section 4.5.3 includes a discussion of the results in the data analysis and shows my findings and implications.

I conclude in Section 4.6.

4.2 Attention-Based View and antecedents of convergence

The Attention-Based View of the firm (ABV) (Ocasio, 1997; Ocasio and Joseph, 2005) is an established theoretical perspective in management research on firm strategic behavior. The ABV theorizes firm strategic behavior as an outcome of the focusing and channelling of management attention. Here, management attention is defined as “*the noticing, encoding, interpreting, and focusing of time and effort by organizational decision makers on both problems and solutions*” (Ocasio, 1997). The theoretical tenet that the focus of management attention leads to firm strategic behaviour has received systematic empirical proof in a variety of contexts and for several strategic behaviours including: responses to institutional change (Nigam and Ocasio, 2010; Ocasio and Radoynovska, 2016); multinational strategy (Bouquet and Birkinshaw, 2008); technology strategy (Eggers and Kaplan, 2009); strategic adaptation (Joseph and Ocasio, 2012), and corporate governance (Tuggle et al., 2010).

Elaborating on its core tenet, the ABV further posits that the focus of management attention is situated in the firm’s context and that this includes the environmental stimuli for decision-making, the interactions among participants in the context, and the embodiment of issues and answers in cultural symbols, artefacts, and narratives (Ocasio, 1997). It follows that, if managers and decision makers are subject to a context that shares similarities, they are also likely to focus their attention towards similar issues, hence their firms are likely to exhibit similar strategic behaviors.

In the systemic risk literature, systemic risk has been usually associated to co-movements between components within a (e.g. financial) system (Billio et al., 2012; Bushman et al., 2017). Based on this definition of systemic risk, the ABV provides a theoretical framework to understand and anticipate strategic co-behaviours across firms and industries, through the lens of the convergence, or divergence, of the constituent elements of the contexts in which management attention is situated.

To operationalize this theoretical insight borrowed from management theory, the first challenge is the identification of salient descriptors of the context in which risk managers’ attention is situated. There are several descriptors that might qualify – e.g., company’s website, risk management presentation, industry white paper, etc. Yet, an important extra feature of such descriptors is their availability and comparability for individual firms and across time. Here I propose that a company’s 10-K report – the document that all listed companies have to submit yearly to the US Securities and Exchange Commission (SEC) – is a salient descriptor of such context and allows detailed comparisons among firms cross-

sectionally and over-time. In fact, 10-Ks have been used as data sources for describing managers and firms behaviours in several academic publications (e.g., Bao and Datta, 2014; Bushman et al., 2017; Cazier et al., 2021; Dutt and Joseph, 2019; Guo et al., 2017). The 10-K reports are structured in several sections which, jointly, describes a company's business, the risks it faces, and the operating and financial results for the fiscal year. To capture the context of risk managers' situated attention, I focus on the 10-K's Risk Factors section, which includes information about the most significant risks that apply to the company or to its assets. It is worth noting that though the 10-K forms are submitted to regulators for regulatory purposes, they are used by us to understand the antecedents of attention, and the analysis of the data can in turn inform regulatory responses. The data have limitations, e.g., the dataset is rigid and formal. However, we choose 10-K filings for our research because: (a) every US listed company has it in each year, (b) consistent format, (c) there is a section specific on Risk Factors for us to study systemic risk, (d) and we will show later that such rigidity is in itself a type of systemic risk.

Building on these premises, I argue that the extent to which company 10-K's Risk Factor sections converge among each other is a predictor of convergence in the focus of risk manager attention, hence of more similar firm strategic behaviours in response to the risk environment. Overall, I contend that convergence among the content of companies' 10-K Risk Factor reports is an indicator of systemic risk.

Being able to measure convergence in risk manager attention could give policy makers and risk professionals a grounded and innovative means for elucidating systemic risk. In Section 4.3, I offer a methodology to operationalize these theoretical insights, by introducing a text-based metric of convergence between 10-K's Risk Factors sections.

4.2.1 Antecedents of convergence in risk managers' attention

The extent of convergence among company' 10-K Risk Factors reports can signal an increase or decrease in systemic risk, but does not in itself specify the drivers of such a change. Hence, such a signal might still leave risk professionals and policy makers oblivious as per what might be the suitable prevention or mitigation responses. To tackle this point I start off from the consideration that I should nuance overall convergence indicators by interrogating the specific narratives that risk managers are exposed to.

Risk managers do not operate in a vacuum and they are exposed to considerations around a company's overall strategy. The broader management literature (e.g., Grant, 2021; Porter, 1985, 1986) can help identify some specific dimensions of the context in which risk managers' attention is situated. In what follows I leverage that literature and its insights to identify a taxonomy of general issues to which risk managers are exposed.

First, the strategic management literature has highlighted that in developing their strategy, firms should look at the trends in the macro environment (Drucker, 1995; Grant, 2021) and describe scenarios that a firm is likely to face (Cornelius et al., 2005). The common trends are not limited to specific industries, and apply to the whole economy, e.g., increased attention to customer services, operation of firms, products, asset management, gains and loss of investment, etc. In the context of risk management, I argue that firm's business narrative around common trends is an important dimension that risk managers consider.

Second, Porter (1985, 1986) and the substantial literature spurred from his seminal books on competitive strategy and competitive advantage have identified some factors – often referred as Porter's Five forces – that affect an industries' profitability; namely, suppliers' bargaining power; clients' bargaining power; intensity of competition; threat of new entrants; and threat of substitution. Here, threat of substitution occurs when companies of one industry have to compete with industries producing substitute products or services. For example, insurers sell financial derivatives that are traditionally sold by banks. While narratives around, e.g., competition and buyers/clients' bargaining power, are related to dynamics within a firm's legacy industry, narratives around substitutes are more likely to be related to cross-industry dynamics and that, as such, attracts the attention of risk managers.

Finally, the management literature has also revealed that often, strategic choices find their antecedents in management fashions (Abrahamson, 1996) and institutional legitimacy considerations (Suchman, 1995). Here I note that for studying the antecedents of systemic risk, fads that are relevant for risk managers should be considered, too. Some of the risk management simply follows fashion rather than discloses real need (Milos et al., 2008), which reduces the usefulness of risk management (Power, 2009).

With the goal of elucidating antecedents of systemic risk, my insights thus far suggest that risk managers are exposed to consideration about trends, substitution and risk management fashions and that, as such, being able to identify and measure such components of the context in which risk manager attention is situated is a salient indication of the antecedents of systemic risk.

In Sections 4.4, I present in detail the construction and operationalization of quantitative measures of those antecedents based on the 10-K reports, and then illustrate examples of how to interpret the results to elucidate the antecedents of systemic risk.

4.2.2 Implications for risk management

My theoretical framework and analysis has outright implications for risk professionals and policy makers. Table 4.1 shows of possible prevailing antecedent of systemic risk, relevant implications, and risk management responses.

If the first antecedent – managers’ attention focusing on common general trends – prevails, there will be both cross-industry and economy-wide risk. Managers’ attention to common issues in the environment, e.g., stock market performance, supply chain operation, customer service, will be the main cause of the convergence. Systemic risk caused by this antecedent is difficult to fully control (i.e. reduce a risk event’s probability of happening) because it is related to system-wide challenges; Nonetheless, it can be to an extent mitigated (i.e. reduces the impact of a risk event) (Hopkin, 2018). At individual firm level, firms can transfer and diversify the risk. For example, both insurers and banks have investments in equities and will therefore suffer in the scenario of a stock market crash. They prepare for such risk by transferring the risk to other entities, e.g., buying reinsurance and financial derivatives. Firms may also face risk from supply chain failure, and they can manage this risk by diversifying suppliers. At macro level, regulators help individual firms control the risk by setting risk limits, and guiding them to prepare for emerging system-wide risks.

If the second antecedent – managers’ attention focusing on the adjacent industry and threat of substitution – is the main driver of the convergence, there will be cross-industry risk. If this is the case, insurers and banks discuss issues traditionally from the other industry because of potential substitution, e.g. on firms’ selling products or adopting strategies traditionally associated with the other industry. Increased attention on this issue by a firm, is an indicator of the firm engaging in such practices (or indeed being exposed to others doing so). To control and mitigate such systemic risk, regulators build firewalls to limit the blurring of industry boundaries. For instance, insurers are regulated to have limited freedom to sell banking products, because insurers’ involvement in banking business, e.g., selling structured credit products, is believed to be the main cause of the failure of AIG in 2008 (Cummins and Weiss, 2009).

If the third antecedent – risk managers’ exposure to similar risk management fashions – prevails, then both idiosyncratic and economy-wide risks arise. Managers’ attention may be led to “fashionable” topics, e.g., implementing enterprise risk management (ERM) frameworks without sufficiently understanding risk appetite and reflecting on non-quantifiable uncertainties (Power, 2009). Consequently, the danger of groupthink occurs with substantial risks (idiosyncratic or otherwise) ignored and risk disclosure becoming less useful in reflecting and anticipating the risks the firms are exposed to. The ignored risks may accumulate, as they are not dealt with in a timely way, and finally burst out together, causing system-wide effects. Another detrimental consequence of groupthink is the ignorance of firms’ unique competitive advantages, together with idiosyncratic risk. Groupthink may also lead to similar actions adopted when facing the same event, which further exacerbates systemic risk. Previous studies show that corporate governance, especially the role of risk

professionals, has a big impact on the performance of firms during crises (Aebi et al., 2012; Hunjra et al., 2020). Hence, systemic risk derived from this antecedent can be controlled and mitigated by corporate governance, e.g., ensuring diversity in board members' and risk managers' background, expertise, and perspectives.

Prevailing antecedent	Implications	Risk management response
<u>Common Trends</u> : attention on common general trends in the macro environment.	Cross-industry and economy-wide risk. Common risk exposures across industries.	Risk transfer; risk diversification; regulation on risk limit and preparedness.
<u>Substitution</u> : attention on the adjacent industry and threat of substitution.	Cross-industry risk. Firms in adjacent industries facing common threat because of increased similarity in their products and strategies.	Regulation on firewalls to limit the blurring of industry boundaries.
<u>Fashion</u> : exposure to similar risk management fashions.	Both idiosyncratic and economy-wide risk. Substantial risks (idiosyncratic or otherwise) are ignored and risk disclosure becomes less useful in timely reflection on / anticipation of potential threats. The ignored risk accumulate and burst out together, causing system disorder.	Corporate governance, diversification in boards, particularly in the professional group of risk managers.

Table 4.1 Antecedents of convergence, implications and risk management response

4.3 Measuring inter-industry convergence

4.3.1 Data

My dataset contains 10-K submissions to the SEC of all the 214 banks, 94 insurers and 223 pharmaceuticals listed on the New York Stock Exchange from 2006 to 2018. I focus on the two most substantial sections of those reports: the Business and Risk Factors sections. Using 10-Ks as a source of systematic information is increasingly common in management and finance research (Bushman et al., 2017; Dutt and Joseph, 2019; Nadkarni and Chen, 2014). I collect all the 10-K reports manually from Filings Expert. I focus on the insurance companies classified as insurers or banks (based on the North American Industry Classification System, NAICS, codes) as these are the industries across which I aim to measure convergence in

attention. In addition, I use text data from pharmaceutical's 10-K reports, as a benchmark industry, that is not closely related to either insurance or banking – data from pharmaceuticals will be used in Section 4.4 for constructing a measure of risk management fashions.

Thus my text data cover $T = 13$ years and $I = 531$ companies. Define the sets $\mathcal{I} = \{1, \dots, I\} = \mathcal{I}_{IN} \cup \mathcal{I}_{BK} \cup \mathcal{I}_{PH}$, where $i \in \mathcal{I}_{IN}$ if the i -th company is an insurer, $i \in \mathcal{I}_{BK}$ if it is a bank, and $i \in \mathcal{I}_{PH}$ if it is a pharmaceutical. Thus each 10-K report in my data corresponds to a pair (i, t) , $i \in \mathcal{I}$, $t \in \mathcal{T}$. Let J_{it} be equal to one if there is a document for firm i in year t and zero otherwise. The numbers of documents for in year t for each industry are then denoted as

$$n_{A,t} = \sum_{j \in \mathcal{I}_A} J_{j,t}, \quad A \in \{IN, BK, PH\}.$$

The total number of firms in each industry and year can be found in Table 4.2.

Table 4.2 Total number of firms for each industry and year.

	Insurers	Banks	Pharmas
2006	53	173	94
2007	55	171	99
2008	54	177	105
2009	54	182	113
2010	76	191	116
2011	82	194	118
2012	61	193	110
2013	84	202	103
2014	90	204	102
2015	89	200	94
2016	84	189	100
2017	60	171	92
2018	74	131	91

Before any text analysis is carried out, some standard pre-processing steps are applied: (a) Stop words are those words most likely to appear in all documents in the corpus (e.g., “a,” “the,” “of”, etc.), and they carry little semantic meaning (Bengfort et al., 2018, p.65). Stop words are removed, which help improve the performance as there are fewer and only meaningful tokens left (Bengfort et al., 2018, p.72-74). (b) Bigrams, possible contiguous subsequences of two words, are constructed. For example, ‘real estate’ is understood as a one term to be meaningful when the two words appear together (Bengfort et al., 2018, p.132-145). (c) All the words and bigrams are lemmatized, i.e. I remove inflectional endings and return the base or dictionary form of a word (Stanford, 2009b).

Note that in the current section I only consider banks' and insurers' reports and only their Risk Factors sections. Report from the Business sections and for pharmaceuticals will be used in Section 4.4 to construct drivers of convergence to be used in regression analysis.

4.3.2 Construction of the convergence measure

My purpose in this section is to construct a measure of convergence, that reflects the similarity of an individual bank's (resp. insurer's) Risk Factors section in their 10-K report, to the insurance (resp. banking) industry as a whole. This measure is meant to represent managers' convergence of attention. Constructing such a measure entails two distinct steps. First, each individual document (Risk Factor sections of 10-K reports), needs to be converted to a vector representation. Second, a measure of similarity is employed to calculate inter-industry convergence, based on those vectors.

To convert each document into a vector, I use word embeddings constructed by Doc2vec, which is an unsupervised algorithm that learns fixed-length feature representations from documents of varying length (Le and Mikolov, 2014). Word embeddings are well suited for mapping documents' semantic content onto a vector space, by explicitly considering the context in which individual words occur. Such context is not captured in standard approaches that are based purely on word frequencies e.g. frequency, one-hot, and TFIDF encoding (Bengfort et al., 2018; Kusner et al., 2015, p.65-66).

To derive word embeddings from our data, I use Gensim's Doc2vec class in Python. I choose Doc2vec over alternatives such as Word2Vec embeddings (Mikolov et al., 2013b), because Doc2vec has been found to outperform, e.g., simple averaging of Word2Vec vectors, in terms of the error rates of an information retrieval task (Le and Mikolov, 2014). The Doc2vec algorithm represents each document by a dense vector which is trained to predict words in the document, using neural network.

Given the representation of each document by a vector, I measure the convergence in management attention across organizations in distinct industries using silhouette values (Rousseeuw, 1987). This is a special case of the micro-macro across-class convergence proposed in Section 2.3.1 with only two classes. The silhouette value is a measure, used in unsupervised learning, of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It is calculated using the mean intra-cluster (here: intra-industry) distance and the mean nearest-cluster (here: inter-industry) distance for each observation. A low value indicates that the object is poorly matched to its own cluster and well matched to neighboring clusters, indicating, in my context, high convergence.

I define the vector representing each Risk Factors document for insurers and banks by D_{it} , $i \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK}$, $t \in \mathcal{T}$. To calculate convergence for insurers, for each document with

$i \in \mathcal{I}_{IN}$ and $J_{it} = 1$, I define the quantities:

$$a_{it} = \frac{1}{n_{IN,t} - 1} \sum_{j \in \mathcal{I}_{IN} \setminus i} J_{jt} \cdot d(D_{it}, D_{jt}),$$

$$b_{it} = \frac{1}{n_{BK,t}} \sum_{j \in \mathcal{I}_{BK}} J_{jt} \cdot d(D_{it}, D_{jt}),$$

where d represents the Euclidean distance between document vectors. Here, a_{it} represents the average dissimilarity of an insurer's Risk Factors section to other insurers' reports (across years) while b_{it} reflects the average dissimilarity to banks' reports. For $i \in \mathcal{I}_{BK}$ I define the quantities a_{it}, b_{it} analogously. Then, my measure of inter-industry convergence is given by the *negative silhouette value* defined as

$$Convergence_{it} = -\frac{b_{it} - a_{it}}{\max\{a_{it}, b_{it}\}}, \quad i \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK}. \quad (4.1)$$

While the silhouette value in clustering analysis represents the cohesion of clusters, its negative, used here, reflects the extent to which a document is close to documents from a different industry, compared to the baseline of the document's closeness to other documents from its own industry.

4.3.3 Results

In Figure 4.1, I present the average and 95% confidence interval of the yearly value of $Convergence_{it}$ for insurers and banks. Larger value of $Convergence_{it}$ imply more convergence in risk management disclosure. There is a generally decreasing trend of $Convergence_{it}$ in the insurance industry over the years after 2008, while the trend of banking industry is increasing.

It is worth noting that the trend of the insurance industry is not consistent with the trend in Figure 3.2. This implies that the trend in convergence of the insurance industry is not robust, and we will discuss it in more detail in Chapter 2. This conclusion does not contradict the phenomenological evidence of convergence in the financial industry, if we make distinction between “convergence” and the “trend of convergence”. The convergence of the insurance industry towards the banking industry is always there, and it is higher than that of the banking industry. However, its trend is not robustly monotonically increasing, which might be lessened by factors such as Subprime Mortgage Crisis.

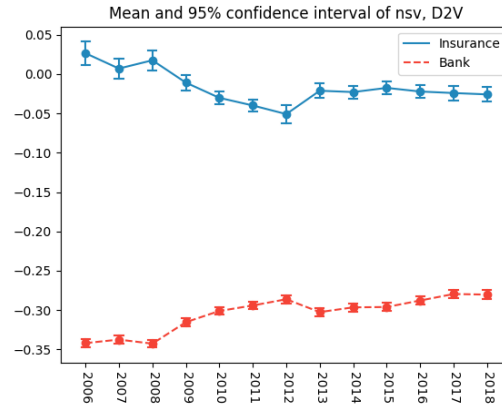


Fig. 4.1 Means and 95% confidence intervals of convergence of risk disclosure in each year by industry.

4.4 Construction of antecedents of convergence

The calculation of the negative silhouette value on Doc2vec representations of Risk Factor sections in Equation (4.1) enabled us to construct a convergence measure of management attention reflected by risk disclosure. Now I turn my attention to construct the antecedents. I first perform a topic modelling analysis, in order to identify general themes from the analysed text documents. This helps me to derive interpretable descriptors of the composition of individual documents, which in turn will be used to construct antecedents of convergence in Section 4.2.

4.4.1 Topic modelling of 10-K Business sections

A topic model is a type of statistical model for discovering a set of topics that are shared by a collection of documents. The most common topic model currently in use is the Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003). Examples of application include Bao and Datta (2014); Bellstam et al. (2021); Taeuscher et al. (2021). LDA generates summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers, for each document, the distribution of its word content over topics.

LDA is a generative model, with the following underlying assumptions. A fixed number of topics, K exist. Each document is assumed to be randomly constructed, by choosing its composition by topic according to a Multinomial distribution with Dirichlet prior. Furthermore, for each topic a Multinomial distribution over the vocabulary of size N is assumed, again with Dirichlet prior, giving the probability of any word in the vocabulary belonging to this topic.

I run the LDA model on all the Business section from 10-K reports of all the three industries in all years. I use Business section rather than Risk Factors section to (a) reflect the strategy of the firms, and (b) avoid endogeneity, as the variables constructed based on LDA will be used in regression later. The number of topics K is a parameter set by the user of LDA. Since I use the topic distribution later to construct antecedents of the convergence, representing the attention on each antecedents, I chose K subjectively, such that the topics generated are distinct and interpretable. As the number of topics increases, the generated topics become less distinct and interpretable (e.g. different topics that have very similar words). I chose the largest topic number that generates the distinct and interpretable topics, so as to extract maximum information. LDA estimates the following quantities:

- a) For each document (i, t) , $i \in \mathcal{I}$, $t \in \mathcal{T}$, $k = 1, \dots, K$, p_{itk} gives the proportion of the topic k in that document.
- b) For each word $n = 1, \dots, N$ in the vocabulary, q_{nk} gives the relative frequency by which the word will appear as part of of topic k .

In Figure 4.2, I show word clouds generated for each of the topics, representing the estimates q_{nk} . I interpret these word clouds as distinct themes:

- Topics 1 and 3 are insurance-specific, referring to insurance liabilities and life policies respectively.
- Topics 2 and 4 are idiosyncratic to banking, discussing loans and deposits, and regulation and capital.
- Finally, topics 5 and 6 do not appear to be particular to any single industry, the first referring to operations, products and customer service; the latter to investments and assets.

In Table 4.3, I present the mean distribution of topics, for documents within each industry; that is, I show the average of p_{itk} over $t \in \mathcal{T}$ and over i in each of $\mathcal{I}_{IN}$, $\mathcal{I}_{BK}$. Frequencies higher than 5% are highlighted. I see that the highest frequencies in each row of Table 4.3 are consistent with the way that I have interpreted Figure 4.2.

4.4.2 Antecedent 1: common trends exposure

The first contextual antecedent of firm strategy discussed above relates to the challenges and opportunities in the industry environment. To quantify this antecedent I introduce the variable *common trends*, denoted as $CommonTrend_{it}$. From the LDA analysis of Section 4.4.1, Figure

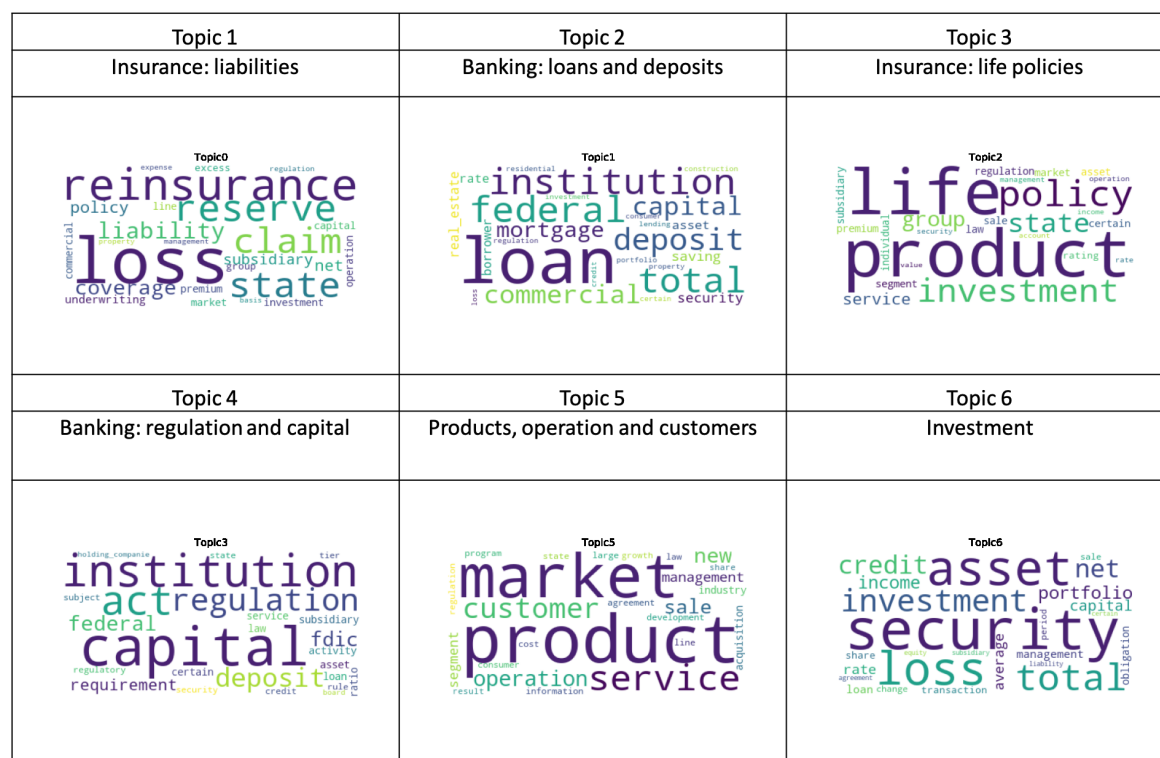


Fig. 4.2 LDA-generated word clouds, representing the distribution of words within each topic.

Table 4.3 Average percentage distribution of topics, for documents within each industry.

		Insurance	Banking	Pharmaceutical
Topic 1	Insurance: liabilities	50.83	0.34	0.19
Topic 2	Banking: loans and deposits	3.02	33.33	0.19
Topic 3	Insurance: life policies	23.50	1.27	0.31
Topic 4	Banking: regulation and capital	4.39	53.59	0.27
Topic 5	Products, operation and customers	10.07	4.54	12.60
Topic 6	Investment	7.91	6.82	1.36

4.2, I identified two topics that reflects aspects of the environment that cut across industries, specifically Topics 5 ('Products, operations, and customers') and 6 ('Investment'). Hence I define:

$$CommonTrend_{it} = p_{it5} + p_{it6}, \quad i \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK}.$$

4.4.3 Antecedent 2: threat of substitution

The second contextual antecedent of firm strategy relates to the potential for substitution and new industry entry. I measure the relevance of this antecedent for specific firms, via the variable *threat of substitution*, denoted as $Substitution_{it}$ and defined as follows. From the LDA analysis of Section 4.4.1, I have seen that there are insurance-specific and bank-specific topics. The variable $Substitution_{it}$ reflects the extent to which a firm (e.g. an insurer) places attention on topics specific to the industry it does not belong to (e.g. banking), measured by the percentage of the firm's 10-K Business section spent on those topics. Hence, for an insurer, $i \in \mathcal{I}_{IN}$, I have that:

$$Substitution_{it} = \begin{cases} p_{it1} + p_{it3}, & i \in \mathcal{I}_{IN}, \\ p_{it2} + p_{it4}, & i \in \mathcal{I}_{BK}. \end{cases}$$

4.4.4 Antecedent 3: management fashions

Finally, I identified as the third contextual antecedent of firm strategy, related to industry convergence, the level of attention to management fashions. I quantify such attention, via the variable $Fashion_{it}$.

To construct this variable, there are two steps. In the first step, I measure the negative silhouette value between the Risk Factors sections of 10-K reports, but now with clusters formed by financial firms (insurers and banks combined) and pharmaceuticals. Specifically, I let

$$SVFP_{it} = -\frac{\tilde{b}_{it} - \tilde{a}_{it}}{\max\{\tilde{a}_{it}, \tilde{b}_{it}\}}, \quad i \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK},$$

where,

$$\tilde{a}_{it} = \frac{1}{n_{IN,t} + n_{BK,t} - 1} \sum_{j \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK} \setminus i} J_{jt} \cdot d(D_{it}, D_{jt}),$$

$$\tilde{b}_{it} = \frac{1}{n_{PH,t}} \sum_{j \in \mathcal{I}_{PH}} J_{jt} \cdot d(D_{it}, D_{jt}).$$

The logic of such a construction is as follows. The variable $SVFP_{it}$ represents the convergence of attention of risk managers in individual financial firms (insurers or banks) to the pharmaceutical industry. I consider the pharmaceutical industry to be very distinct to the financial serviced industry, with, for example, no opportunities for financial firms entering the pharma industry (and vice verse). Hence, $SVFP_{it}$ reflects the extent to which attention is placed on more general narrative about risk, which cut across the boundaries of industries that are not closely related.

It may be argued that $SVFP_{it}$ and $CommonTrend_{it}$ are correlated, and the two concepts intertwined with each other. To solve this problem, in the second step, I regress $SVFP_{it}$ against $CommonTrend_{it}$ using Ordinary Least Squares regression (OLS), and $Fashion_{it}$ is measured using the *residual* of this regression.

4.4.5 Control variables

In Section 4.5 I will present non-linear regressions of $Convergence_{it}$ on the metrics I just introduced to represent the theoretically derived antecedents of convergence. For this regression I introduce a number of statistical controls. All control variables are defined for $i \in \mathcal{I}_{IN} \cup \mathcal{I}_{BK}$, $t \in \mathcal{T}$.

Year fixed effects. The time $t \in \mathcal{T}$ corresponding to each 10-K is used as a categorical variable to control for trends that are not reflected in the convergence antecedents.

Industry fixed effects. The industry a firm belongs to, i.e. insurance or banking is used as a categorical variable, to control for firms' in different industries potentially responding differently to convergence drivers.

Inter-industry stock return correlation. I measure Kendall's rank correlation between an individual insurer's (resp. bank's) daily stock return and the return on the S&P 500 Bank Index (resp. Insurance Index). This variable, which I call $StockCorr_{it}$, is meant to control for 10-K reports responding to external shocks that jointly move insurance and banking stocks. For each (i, t) , the value of $StockCorr_{it}$ is calculated on stock market data over the one-year time period before the disclosure date of the corresponding 10-K report. I use Kendall's rank correlation because it is robust to extreme observations (Kendall, 1938). I preferred it over

tail risk measures (e.g., Bushman et al., 2017), as I am not specifically focused on extreme levels of stock co-movement.

Data on the stock price of each company come from the Quandl database. Data for the industry indexes are downloaded from S&P Global Market Intelligence. In regressions, $StockCorr_{it}$ is standardized by industry.

Non-informative language. It may be plausible that the 10-K reports of a firms is similar to reports in a different industries, because of its use of non-informative language in risk disclosures, which is heavy with cliches, that is, standardized expressions carrying little meaning. To control for this effect, I quantify language non-informativeness, by slightly adpating the ‘boilerplate’ language metric of Lang and Stice-Lawrence (2015). The resulting variable $Boiler_{it}$ is calculated by the same process as in Section 3.3.4. In regressions $Boiler_{it}$ is standardized by industry.

4.5 Explaining convergence of risk disclosures in terms of contextual antecedents

In this section I empirically investigate the extent to which the contextual antecedents of firm strategy drive convergence of attention. Empirically, I carry out regression models of the form $E[Convergence_{it} | \mathbf{X}_{it}] = g(\mathbf{X}_{it})$, where g is a (non-linear) regression function and $\mathbf{X}_{it}$ includes the covariates and controls developed in Section 4.4.

4.5.1 Model selection

First, I select the best regression model from a set of standard regression modelling approaches. In particular, the five models I consider are Lasso, Ridge, Elastic Net, Random Forest, and Gradient Boosting Tree regressions. All regressions are carried out in Python, using the scikit-learn package (?). To evaluate the models, I use cross-validation with 20 folds. As I have far more banks than insurers in my sample, I make sure all folds are stratified and balanced, to have the same proportion of insurers and banks in each fold. I run each model for different parameter settings and report the performance results using Mean Squared Errors (MSE), averaged over all folds. These prediction errors are reported in Table 4.4, for different hyperparameter choices, in an approach similar to Agarwal et al. (2019) (all other hyperparameter values are set at their default values).

In the following analysis, I use the results from the Random Forest regression model with 100 trees due to its lowest reported MSE. The superior predictive performance compared to linear models, indicates the presence of important non-linear effects and variable interactions.

To disentangle these relationships, I carry out a detailed discussion of variable importance and sensitivity analysis in Section 4.5.2.

Table 4.4 Results of the performance of different regression models with different parameters

Regression model	Parameter	Param value	Average MSE
Lasso	Constant that multiplies the L1 term	0.01	0.5398
		0.1	0.5866
		0.5	0.8923
		1	0.8954
		10	0.8954
		100	0.8954
Ridge	Regularization strength	0.01	0.5410
		0.1	0.5410
		0.5	0.5410
		1	0.5410
		10	0.5413
		100	0.5443
ElasticNet	Penalty terms	0.01	0.5407
		0.1	0.5679
		0.5	0.7204
		1	0.8932
		10	0.8954
		100	0.8954
RandomForest	The number of trees in the forest	10	0.4345
		50	0.3687
		100	0.3550
		1000	0.3604
		10000	0.3622
GradientBoosting	The number of boosting stages to perform	10	0.6117
		50	0.4395
		100	0.4203
		1000	0.3847
		10000	0.4186

4.5.2 Variable importance and sensitivity

Accumulated Local Effects

I visualize the results of the chosen regression model, via Accumulated Local Effects (ALE) plots and variable importance measures. ALEs, proposed by Apley and Zhu (2020), describe how features influence the prediction of a machine learning model on average. ALE plots shows how the model predictions change in a small ‘window’ of the feature around a certain

grid value for data instances in that window. ALEs can be efficiently calculated and are not distorted by correlation between features (Molnar, 2020), which affects alternative measures such as Partial Dependence Plots.

The ALE plots for my chosen regression model are displayed in Figure 4.3. The impact of extreme values (bottom and top 5% of feature values) of each variable is not shown (except for *Year*), as these are very noisy. I observe the following:

- For all three the covariates, *CommonTrend*, *Substitution* and *Fashion*, I see a clear positive impact on the response variable *Convergence*, which reflects convergence in attention across industries.
- From observing the scale of the respective ALEs, I note that *Fashion* is the most impactful variable.
- I also see that the control *StockCorr* is positively associated with the response, which reflects my assumption that co-movements of stock prices would have a positive impact on attention convergence, which is controlled for.
- The ALE for *Year* also shows an increasing pattern, reflecting an underlying positive trend in convergence in general.
- Finally, there is no clear pattern for *BoilerPlate*, but the scale of the ALE is much smaller, hence I consider this variable inconsequential.

Feature importance

In addition, I can quantify the relative importance of the different covariates. For this purpose I use two variance importance measures: the impurity-based feature importance and the variance permutation importance (Breiman, 2001). For each of the two methods, the higher the value of the measure is, the more important the feature is. Impurity-based feature importance is specific to tree-based methods and is computed as the (normalized) total reduction of the criterion (e.g., Gini or Mean Squared Error) brought by that feature. A split (i.e., the separation of data into different nodes in a particular tree of the Random Forest) with a large decrease of impurity is considered important, and therefore, variables used for splitting at important splits are also considered important. The impurity is usually measured by the Gini impurity (Ishwaran, 2015). Permutation importance is defined to be the difference between the baseline prediction error of the fitted model and the error arising after randomly permuting a given feature column (Breiman, 2001). Impurity-based feature

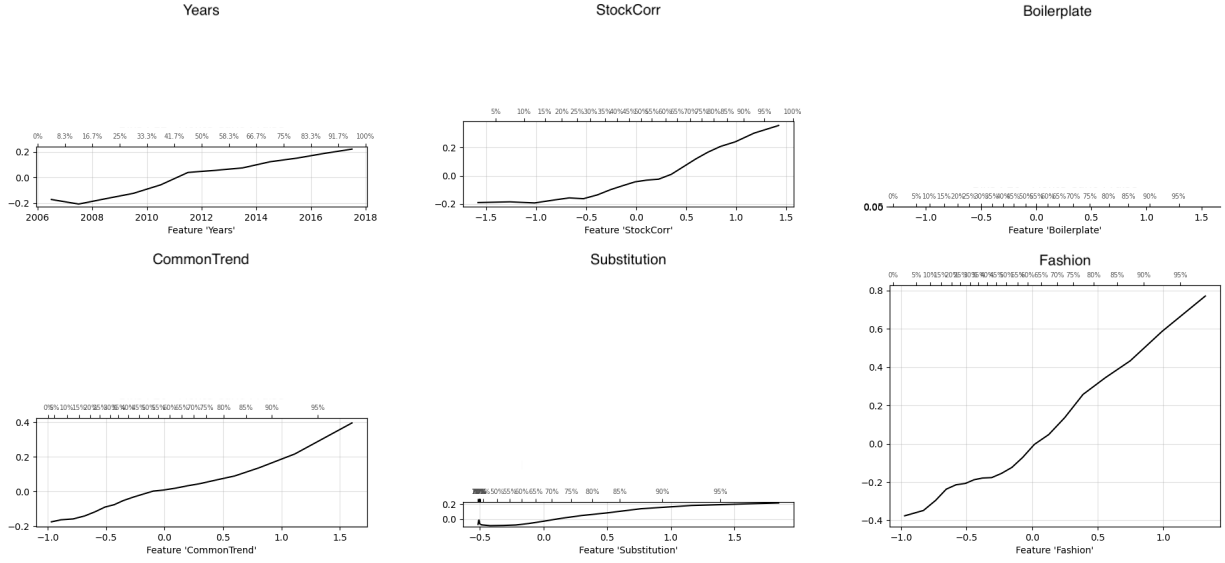


Fig. 4.3 ALE plots for the Random Forest regression model.

All the vertical axes are rescaled to make all the subfigures are comparable.

importance can be misleading for high cardinality features and permutation importance has been shown as an appropriate alternative to solve this problem (Breiman, 2001).

The feature importance and permutation importance of all the drivers of the Random Forest regression model can be found in Figure 4.4. According to both importance measures, the most important feature *Fashion*. The is consistent with the picture provided by ALEs and establishes *Fashion* as the most important driver of convergence.

Sensitivity analysis

Finally, I evaluate the contribution of individual covariates to predictions at different (e.g. high) levels of convergence. I follow the Marginal Attribution by Conditioning on Quantiles (MACQ) framework of Merz et al. (2021) (see also Tsanakas and Millosovich (2016)). For this, I require an additive decomposition of individual predictions, such that for the i -th observation in the sample, with features $\mathbf{x}_i$ and predictions $\hat{g}(\mathbf{x}_i)$, I have

$$\hat{g}(\mathbf{x}_i) = \sum_j \phi_j(\mathbf{x}_i), \quad (4.2)$$

where $\phi_j(\mathbf{x}_i)$ represents the contribution of the j -th covariate to the prediction of the i -th observation. In Merz et al. (2021) such a decomposition was derived by a quadratic approximation to smooth prediction functions of deep learning models. Here, given the non-smoothness of

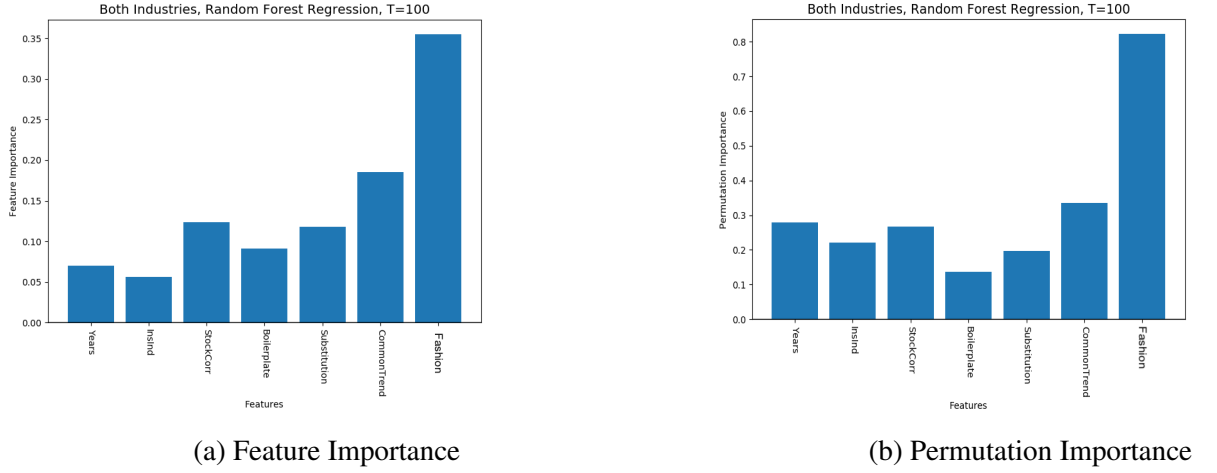


Fig. 4.4 Importance of each antecedents.

(a) shows impurity-based feature importance, while (b) shows permutation importance.

the fitted Random Forest's prediction function, I instead carry out decomposition in Equation (4.2) via Shapley values.

The Shapley value for a feature-observation combination is derived from the marginal contribution of the feature to individual predictions. Such marginal contributions are averaged across 'coalitions' of features to which the one under focus is added. Thus, differently to feature importance measures, Shapley values quantify the contribution of each covariate at the *individual* observation level. Shapley values were originally defined in the context of co-operative game theory (Shapley, 1953) and have become popular for interpreting the predictions of machine learning following the SHAP framework of Lundberg and Lee (2017a,b) – for a technical overview see Aas et al. (2021).

I plot in Figure 4.5 (top) for each of the features *CommonTrend*, *Substitution*, and *Fashion*, the attributions $\phi_j(\mathbf{x}_i)$ on the vertical axis, against the quantile level u_i of the prediction $\hat{g}(\mathbf{x}_i)$, that is, $\hat{g}(\mathbf{x}_i) = \hat{F}^{-1}(u_i)$, where $\hat{F}$ is the empirical distribution of predictions. The red curves show non-parametric estimates of the function $u \mapsto E[\phi_j(\mathbf{X}) \mid \hat{g}(\mathbf{X}) = \hat{F}^{-1}(u)]$, with the expectation taken over the empirical measure. Finally the color of the plotted points represents the value of the feature examined. If we focus specifically on the right of each plot in Figure 4.5, e.g. $u \in [0.8, 1)$, we consider those firm-years for which the highest convergence is observed. We can see that for such companies, which are most relevant for considerations of systemic risk, the contribution of *Fashion* is by far the most important one.

In Figure 4.5 (bottom) I plot the attributions $\phi_j(\mathbf{x}_i)$ against the feature values $x_{i,j}$ – the colours now represent the response quantile level u_i .¹ I observe that the relationships between covariates $x_{i,j}$ and attributions $\phi_j(\mathbf{x}_i)$ are clearly increasing, with the points at the top right typically displaying high values of the response variable, convergence. The steeper increase of in the plot for *Fashion* once again confirms the dominant effect of this variable.

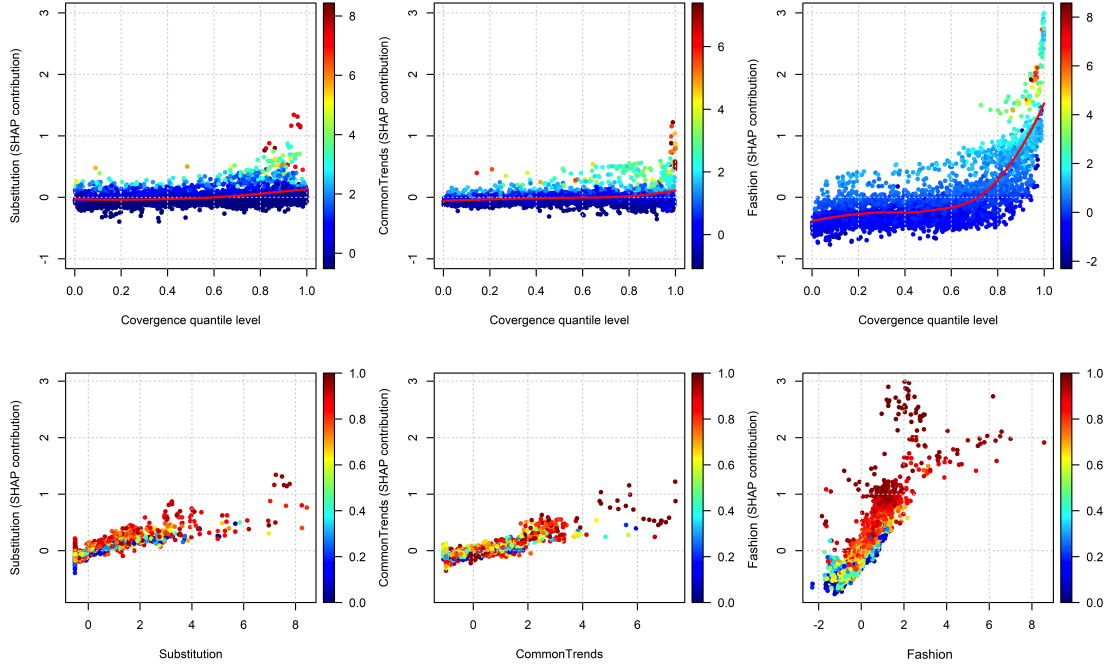


Fig. 4.5 Results of sensitivity analysis based on Shapley values.

Top: feature contributions to predicted value, against quantile level of predictions; color represents the feature value. Bottom: feature contributions to predicted value, against feature values; the colors represent the quantile level of predictions.

4.5.3 Implication of empirical results

My analysis showed that *Convergence* is increasing in all three variables representing the antecedents of attention that I identified. This lends empirical support to my theoretical framework developed in Section 4.2. Furthermore, *Convergence* is increasing in time, implying that over the years, all the antecedents of convergence in management attention of insurers and banks is increasing, indicating the potential for systemic risks arising in the future.

¹Note that for a linear model of the form $\hat{g}(\mathbf{x}_i) = \sum_j \beta_j x_{i,j}$, Shapley values produce attributions $\phi_j(\mathbf{x}_i) = \beta_j x_{i,j}$, thus reproducing the modelled linear effects.

I have found that the most important antecedent of convergence in the attention of risk managers is management fashions – this is consistent for all three feature importance measures that I used in my analysis. In summary, if the semantic similarity of 10-K reports between insurers and banks is largely explained by the similarity of firms in each of those industries to pharmaceuticals, then I may conclude that the main driver of convergence in attention has less to do with the macro or competitive environment and more with the tendency of firms to speak in a language that spans unrelated industries, thus potentially not reflecting idiosyncratic threats. Hence my empirical analysis raises the danger of groupthink as most pressing, driving firms' attention and thus resources to similar issues, with substantial risks (idiosyncratic or otherwise) ignored. As a corollary, literal reading of risk management disclosures becomes less useful in detecting the risk specific to individual firms. What's worse, this can lead to herding behavior, whereby similar decision frames are adopted, with the resulting synchronized resonance potentially exacerbating external shocks (Danielsson and Shin, 2003). Overall, systemic risk is formed by the confluence of those two effects.

4.6 Concluding remarks

The contribution of this paper is fourfold. First, I theoretically connect systemic risk with the context of decision making, through the lens of the Attention-Based View. Second, building on strategic management theory, I establish a framework for antecedents of convergence in attention, and therefore, systemic risk. The strategic antecedents include common trends in the macro environment, threat of substitution, and management fashions. Third, I operationalize the theoretical framework using text data from risk disclosures and create measures of convergence in attention and its antecedents. Finally, based on the theoretical framework and my findings, I highlight the antecedent (and hence the “variant”) of inter-industry systemic risk. Among the three antecedents, management fashion has the dominant explanatory power, in other words, the semantic across-industry similarity between reports is explained mostly by a shared lexicon. Its relationship with systemic risk is the ignorance of substantial risks (idiosyncratic or otherwise) and herding behavior. The importance of management fashion demonstrates a different type of systemic risk.

I see two issues arising from my discussion, which require further research.

First, the implications for systemic risk of risk management fashion being dominant in explaining convergence could be ambivalent. On the one hand, the fashion may lead to the ignorance of idiosyncratic threats and herding behavior (Section 4.5.3), which increase systemic risk. On the other hand, different industries using the same language to manage risk could make it easier to communicate across industry boundaries, which may in turn

(a) reduce the regulatory gap across industries (the regulatory gap between insurance and banking industries is believed to be the main cause of the 2008 Crisis (Sjostrom, 2009)); (b) allow the earlier detection of economy-wide risks. Such an effect would actually reduce systemic risk. Hence, understanding how these two effects interrelate, and which one has dominant implications for systemic risk, is a topic for future research.

Second, the association of convergence in attention with groupthink potentially reflects a homogeneity of stakeholders like risk managers, who may have similar education backgrounds, use standard risk management tools, and share a prototypical mindset that becomes widely spread through digital technology (e.g., Chapter 3). A common suggested remedy for groupthink is corporate governance (Howard, 2010; O'Connor, 2002). Specifically, one may conclude that the educational, professional and demographic backgrounds of board members and risk managers should be diverse. Senior management diversity is believed to foster strategic change and firm performance (del Carmen et al., 2019). In my context of risk management, to assess the risks that, e.g., a life insurance company exposed to, risk managers need to be able to engage with issues of medicine, law, economics, marketing, etc., beyond a narrow focus on modelling mortality risk and asset values. Nonetheless, to draw this conclusion, the relationship between my measure of risk management fashion and the (lack of) diversity in boards and risk management functions needs to be more closely (and empirically) interrogated.

Chapter 5

Conclusion

This thesis studies the contextual antecedents of inter-industry convergence of management attention and assesses the implication for systemic risk. The Attention-Based View provides a theoretical lens to achieve a more nuanced understanding of the convergence and its antecedents. Systemic risk is the consequence of convergence, and understanding and measuring it are the main purposes of this thesis. A richer conceptualization of systemic risk, via the Attention-Based View and consideration of the antecedents of convergence in management attention, is useful for risk managers and policymakers to predict and manage it.

I use text data to reflect management attention, and develop new measures of text similarity at different aggregation levels. Text data are used to reflect the attention and cognition of managers. They give us information on the three antecedents of the convergence in attention that are considered in the thesis, namely attention to common general trends in the macro environment, attention to the adjacent industry and threat of substitution, and exposure to similar risk management fashions. There are several other text data sources that could potentially be used for this task, e.g., letter to shareholders, conference record of annual meeting with CEO, newspaper articles, etc. It is possible that such sources may also give more direct access to managers' attention and strategic thinking. However, the reasons that I use 10-K forms for our research include: (a) 10-K forms are required by the regulators, so every US listed company submit 10-K forms in each year. Other data sources, e.g., newspaper articles, may leave the sample very unbalanced because some companies, especially small ones, have few chances to appear in newspaper articles, while large companies appear with high frequency. (b) The format of the 10-K forms is consistent across companies and industries, making them comparable. This makes them more amenable for the type of quantitative analysis I pursue in the thesis. (c) 10-K forms have a specific section on Risk Factors, which reflects managers' attention on risk, while other sources are

more general and may not necessarily cover risk. As all datasets have limitations, the biggest cons of 10-K forms is that they may not fully reflect the true attention of the managers, as the forms are submitted to regulators and are visible to the public. The managers will hide their true thinking if they are concerned that revealing it may harm the reputation of and confidence in their company. However, to some extent, all the datasets I mentioned above share the same problem.

My thesis contributes to the scholarship on management attention, by offering a new understanding of the contextual antecedents of management attention, and how they drive convergence in management attention. I establish a framework for antecedents of convergence in attention, and find management fashions have the dominant explanatory power of the convergence. This implies that bearing institutional legitimacy considerations in their mind, risk managers simply follow fashion rather than disclose candidly their risks. This conclusion is contingent on the data that I use. In future research, we may check whether this conclusion holds for other datasets. If not, then the dominance of management fashions is probably derived from the lack of informativeness of 10-K forms rather than convergence of managers' attention itself, which is an important point in its own right.

I inform scholarship in digitalization by conceptualizing digitalization as a pervasive technology with broad scope (Hund, 2020; Turber et al., 2015). I contribute to the ongoing debate on the nature of digitalization and whether it brings organizational similarity (Leonardi and Vaast, 2017; Yoo et al., 2012) rather than silos (Christen et al., 2009; Mabey and Zhao, 2017; Newell et al., 2001). With theoretical analysis and empirical evidence, my conclusion is that due to the pervasiveness nature of digitalization, it brings convergence across organizations in different industries, and across different organizational functions.

Nuancing the convergence antecedents and interrogating the risk disclosure gives policy makers and risk professionals a grounded and innovative means for elucidating systemic risk. I contribute to literature on both systemic risk and informativeness of financial disclosure (Bushman et al., 2017; Power, 2009) by unpacking the multi-faceted nature of systemic risk, and proposing measures of antecedents of systemic risk using financial disclosure. The antecedents of convergence manifest in different ways and have different implications for systemic risk. For example, management fashions, as the most impactful antecedent, exacerbates systemic risk by the ignorance of substantial risks and herding behavior. This challenges the informativeness of risk disclosure, and regulators' over-reliance on the disclosure will leads to delayed detection of risk. The suggestion for risk managers is that when following fashions, e.g., ERM, they should fully understand the key concepts and the underlying logic, and apply them to their own firms bearing firm idiosyncrasy (e.g., firm-specific risk appetites and exposures) in mind.

Methodologically, scholars in all areas may get insights on text similarity measures, which are applied to many area including document classification, automatic question answering, document quality evaluation, etc. (Gomaa et al., 2013). In management studies, text similarity can be used to measure isomorphism, innovation, normative legitimacy, customer agility, experience relevance, etc. In risk management studies, text similarity can be used to reflect interconnectedness. I propose text similarity measures at different aggregation levels, especially the similarity of an individual document against a class, and similarity measures across classes and within a class. Text similarity measures at different aggregation levels provide a more complete description of similarity, especially when studying at macro levels, e.g., when aggregating all the individuals in an industry, country or any other specific types.

There are still some limitations and several open questions unsolved. First, my sample of observations are limited to U.S. listed companies. Moreover, the 10-K data may not truly reflect the managers' thinking, though other data may also have the same problem. In future research, other novel data source can be collected, e.g., by web crawlers, interviews, etc.

The second future research direction is related to the ambivalent impact of management fashion on systemic risk. Convergence in attention driven by management fashions either increases systemic risk by engendering the ignorance of idiosyncratic threats and herding behavior, or it reduces systemic risk by reducing cost of communication among different industries and thus allowing the earlier detection of economy-wide risks. It will be important (if challenging) to find out which effect dominates.

Third, future research may also interrogate the relationship between groupthink and corporate governance (e.g., the diversity of risk managers' background). For example, it will be worthwhile to check the relationship between the level of diversity in the (e.g., educational, cultural) background of board committee members, and convergence in management attention to other industries.

Finally, conclusions based on text similarity may vary with different vectorization methods. Traditionally, the choice of vectorization methods is often determined by the performance of the vectorization in correctly classifying documents, but this criterion may not give useful information if all the methods have good performance. Other criteria, probably qualitative, are needed in future research. For example, we may randomly pick a small sample of the data, and then use manual judgement and discretion, based on relevant theoretical analysis, to decide the best choice of methods.

References

- Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.
- Abbas, A., Zhou, Y., Deng, S., and Zhang, P. (2018). Text analytics to support sense-making in social media: A language-action perspective. *MIS Quarterly*, 42(2).
- Abrahamson, E. (1996). Management fashion. *Academy of management review*, 21(1):254–285.
- Acharya, V. and Dawsonera (2011). *Regulating Wall Street: the Dodd-Frank Act and the new architecture of global finance*. Wiley.
- Acharya, V., Engle, R., and Richardson, M. (2012). Capital shortfall: A new approach to ranking and regulating systemic risks. *American Economic Review*, 102(3):59–64.
- Acharya, V. V., Pedersen, L. H., Philippon, T., and Richardson, M. (2017). Measuring systemic risk. *The review of financial studies*, 30(1):2–47.
- Adamopoulos, P., Ghose, A., and Todri, V. (2018). The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research*, 29(3):612–640.
- Adner, R. and Kapoor, R. (2010). Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strategic Management Journal*, 31(3):306–333.
- Aebi, V., Sabato, G., and Schmid, M. (2012). Risk management, corporate governance, and bank performance in the financial crisis. *Journal of Banking & Finance*, 36(12):3213–3226.
- Agarwal, A., Gupta, A., Kumar, A., and Tamilselvam, S. G. (2019). Learning risk culture of banks using news analytics. *European Journal of Operational Research*, 277(2):770–783.
- Aguinis, H., Edwards, J. R., and Bradley, K. J. (2017). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods*, 20(4):665–685.
- Akstinaite, V., Robinson, G., and Sadler-Smith, E. (2020). Linguistic markers of ceo hubris. *Journal of Business Ethics*, 167(4):687–705.
- Albertini, L. (2009). The investor perspective non-life in the handbook of insurance linked securities.

- Amihud, Y., Mendelson, H., and Pedersen, L. H. (2012). *Market liquidity: asset pricing, risk, and crises*. Cambridge University Press.
- Amit, R. and Zott, C. (2012). Creating value through business model innovation. *MIT Sloan Management Review*, 53(3):41.
- Anandarajan, M., Hill, C., and Nolan, T. (2019). Practical text analytics. *Maximizing the Value of Text Data.(Advances in Analytics and Data Science. Vol. 2.) Springer*, pages 45–59.
- Anckaert, P.-E., Cassiman, D., and Cassiman, B. (2020). Fostering practice-oriented and use-inspired science in biomedical research. *Research Policy*, 49(2):103900.
- Antons, D., Joshi, A. M., and Salge, T. O. (2019). Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*, 45(7):3035–3076.
- Aon (2013). Surety market 2014 outlook.
- Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.
- Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management science*, 57(8):1485–1509.
- Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, pages 265–271.
- Artikis, P. G., Mutenga, S., and Staikouras, S. K. (2008). A practical approach to blend insurance in the banking network. *The Journal of Risk Finance*.
- Arts, S., Hou, J., and Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.
- Ashtor, J. H. (2022). Modeling patent clarity. *Research Policy*, 51(2):104415.
- Avraham, R. (2012). The economics of insurance law - a primer. *Connecticut Insurance Law Journal*, 19:29.
- Bailey, D., Faraj, S., Hinds, P., von Krogh, G., and Leonardi, P. (2019). Special issue of organization science: Emerging technologies and organizing. *Organization Science*, 30(3):642–646.
- Baldwin, C. Y. and Clark, K. B. (2003). Managing in an age of modularity. *Managing in the modular age: Architectures, networks, and organizations*, 149:84–93.
- Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23.
- Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1371–1391.

- Barrett, M., Oborn, E., Orlikowski, W. J., and Yates, J. (2012). Reconfiguring boundary relations: Robotic innovations in pharmacy work. *Organization Science*, 23(5):1448–1466.
- Baum, J. A. and Haveman, H. A. (2020). Editors' comments: the future of organizational theory. *Academy of Management Review*, 45(2):268–272.
- Bellstam, G., Bhagat, S., and Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Management Science*, 67(7):4004–4031.
- Beltratti, A. and Corvino, G. (2008). Why are insurance companies different? the limits of convergence among financial institutions. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33(3):363–388.
- Bengfort, B., Bilbro, R., and Ojeda, T. (2018). *Applied text analysis with python: Enabling language-aware data products with machine learning*. " O'Reilly Media, Inc."
- Bhatia, S., Walasek, L., Slovic, P., and Kunreuther, H. (2021). The more who die, the less we care: Evidence from natural language analysis of online news articles and social media posts. *Risk Analysis*, 41(1):179–203.
- Bhattacharya, M., Inekwe, J. N., and Valenzuela, M. R. (2020). Credit risk and financial integration: An application of network analysis. *International Review of Financial Analysis*, 72:101588.
- Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559.
- BIS (2000). *Principles for the management of credit risk, Banking Supervision and Bank for International Settlements, Basle Committee*. Bank for International Settlements.
- BIS (2003). A glossary of terms used in payments and settlement systems.
- Blanco, C. C., Caro, F., and Corbett, C. J. (2019). Managing safety-related disruptions: Evidence from the us nuclear power industry. *Risk Analysis*, 39(10):2197–2213.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.
- Bloomberg (2021). Bloomberg company profile.
- Boland, Jr, R. J., Lyytinen, K., and Yoo, Y. (2007). Wakes of innovation in project networks: The case of digital 3-d representations in architecture, engineering, and construction. *Organization Science*, 18(4):631–647.
- Boudreau, K. (2010). Open platform strategies and innovation: Granting access vs. devolving control. *Management Science*, 56(10):1849–1872.
- Bouquet, C. and Birkinshaw, J. (2008). Weight versus voice: How foreign subsidiaries gain attention from corporate headquarters. *Academy of Management journal*, 51(3):577–601.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

- Brown, J. S. and Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1):40–57.
- Brown, J. S. and Duguid, P. (2001). Knowledge and organization: A social-practice perspective. *Organization Science*, 12(2):198–213.
- Brownlees, C. and Engle, R. F. (2017). Srisk: A conditional capital shortfall measure of systemic risk. *The Review of Financial Studies*, 30(1):48–79.
- Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Buenstorf, G. and Heinisch, D. P. (2020). When do firms get ideas from hiring phds? *Research Policy*, 49(3):103913.
- Burtch, G., He, Q., Hong, Y., and Lee, D. (2021). How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*.
- Bushman, R. M., Chen, J. V., and Williams, C. D. (2017). Informativeness and timeliness of 10-k text similarity for predicting tail-risk comovement. *Available at SSRN 2983315*.
- Cai, J., Eidam, F., Saunders, A., and Steffen, S. (2018). Syndication, interconnectedness, and systemic risk. *Journal of Financial Stability*, 34:105–120.
- Cao, T., Mu, W., Gou, J., and Peng, L. (2020). A study of risk relevance reasoning based on a context ontology of railway accidents. *Risk analysis*, 40(8):1589–1611.
- Casu, B. and Girardone, C. (2010). Integration and efficiency convergence in eu banking markets. *Omega*, 38(5):260–267.
- Cazier, R. A., McMullin, J. L., and Treu, J. S. (2021). Are lengthy and boilerplate risk factor disclosures inadequate? an examination of judicial and regulatory assessments of risk factor language. *The Accounting Review*, 96(4):131–155.
- CEIOPS (2007). Proposal for a directive of the european parliament and of the council on the taking-up and pursuit of the business of insurance and reinsurance – solvency ii, com (2007) 361 final.
- Cerulli (2021). Outsourcing remains in demand with insurers in europe.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6.
- Chen, H., Cummins, J. D., Viswanathan, K. S., and Weiss, M. A. (2014). Systemic risk and the interconnectedness between banks and insurers: An econometric analysis. *Journal of Risk and Insurance*, 81(3):623–652.

- Chen, K., Li, X., Luo, P., and Zhao, J. L. (2021). News-induced dynamic networks for market signaling: Understanding the impact of news on firm equity value. *Information Systems Research*, 32(2):356–377.
- Chen, K., Zhang, Y., and Fu, X. (2019). International research collaboration: An emerging domain of innovation studies? *Research Policy*, 48(1):149–168.
- Chen, W.-R. and Miller, K. D. (2007). Situational and institutional determinants of firms' r&d search intensity. *Strategic Management Journal*, 28(4):369–381.
- Choudhury, M. and Singh, R. (2021). Identifying factors influencing customer experience in bancassurance: A literature review. *Journal of Commerce & Accounting Research*, 10(2).
- Christen, M., Boulding, W., and Staelin, R. (2009). Optimal market intelligence strategy when management attention is scarce. *Management Science*, 55(4):526–538.
- Christensen, C. M., McDonald, R., Altman, E. J., and Palmer, J. (2016). *Disruptive innovation: Intellectual history and future paths*. Harvard Business School Cambridge, MA.
- Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Clement, M. B. and Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of finance*, 60(1):307–341.
- Clothier, R. A., Greer, D. A., Greer, D. G., and Mehta, A. M. (2015). Risk perception and the public acceptance of drones. *Risk analysis*, 35(6):1167–1183.
- Cohen, M. D., March, J. G., and Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, pages 1–25.
- Cohen, S. L. and Tripsas, M. (2018). Managing technological transitions by building bridges. *Academy of Management Journal*, 61(6):2319–2342.
- Compustat (2021). Compustat.
- Cookson, J. A. and Niessner, M. (2020). Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.
- Cornelius, P., Van de Putte, A., and Romani, M. (2005). Three decades of scenario planning in shell. *California management review*, 48(1):92–109.
- Corritore, M., Goldberg, A., and Srivastava, S. B. (2020). Duality in diversity: How intrapersonal and interpersonal cultural heterogeneity relate to firm performance. *Administrative Science Quarterly*, 65(2):359–394.
- Cowley, A. and Cummins, J. D. (2005). Securitization of life insurance assets and liabilities. *Journal of Risk and Insurance*, 72(2):193–226.
- Criscuolo, P., Dahlander, L., Grohsjean, T., and Salter, A. (2017). Evaluating novelty: The role of panels in the selection of r&d projects. *Academy of Management Journal*, 60(2):433–460.

- Cummins, J. D. and Rubio-Misas, M. (2022). Integration and convergence in efficiency and technology gap of european life insurance markets. *Annals of Operations Research*, pages 1–27.
- Cummins, J. D. and Weiss, M. A. (2009). Convergence of insurance and financial markets: Hybrid and securitized risk-transfer solutions. *Journal of Risk and Insurance*, 76(3):493–545.
- Cyert, R. and March, J. (1963). A behavioral theory of the firm. *Industrial Relations*, 19(3):409.
- Dang, C., Li, Z. F., and Yang, C. (2018). Measuring firm size in empirical corporate finance. *Journal of Banking & Finance*, 86:159–176.
- Danielsson, J., Macrae, R., and Uthemann, A. (2022). Artificial intelligence and systemic risk. *Journal of Banking & Finance*, 140:106290.
- Danielsson, J. and Shin, H. S. (2003). Endogenous risk. *Modern risk management: A history*, pages 297–316.
- Daróczy, Z. (1970). Generalized information functions. *Information and control*, 16(1):36–51.
- D’Aveni, R. A. and MacMillan, I. C. (1990). Crisis and the content of managerial communications: A study of the focus of attention of top managers in surviving and failing firms. *Administrative Science Quarterly*, pages 634–657.
- De Jonghe, O. (2010). Back to the basics in banking? a micro-analysis of banking system stability. *Journal of financial intermediation*, 19(3):387–417.
- Degl’Innocenti, M., Kourtzidis, S. A., Sevic, Z., and Tzeremes, N. G. (2017). Bank productivity growth and convergence in the european union during the financial crisis. *Journal of Banking & Finance*, 75:184–199.
- del Carmen, Triana, M., Richard, O. C., and Su, W. (2019). Gender diversity in senior management, strategic change, and firm performance: Examining the mediating nature of strategic change in high tech firms. *Research Policy*, 48(7):1681–1693.
- Deloitte (2018). Finance 2025: Digital transformation in finance.
- Deng, Y., Zheng, J., Khern-am nuai, W., and Kannan, K. (2021). More than the quantity: The value of editorial reviews for a user-generated content platform. *Management Science*.
- Di, Giuli, A. and Laux, P. A. (2021). The effect of media-linked directors on financing and external governance. *Journal of Financial Economics*.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.
- Drucker, P. (1995). Managing in a time of great change. truman.
- Dungey, M., Luciani, M., and Veredas, D. (2013). Googling sifis. *Systemic risk: Liquidity risk, governance and financial stability*, Forthcoming.

- Dutt, N. and Joseph, J. (2019). Regulatory uncertainty, corporate structure, and strategic agendas: Evidence from the us renewable electricity industry. *Academy of Management Journal*, 62(3):800–827.
- ECB (2020). Financial integration and structure in the euro area.
- Eggers, J. P. and Kaplan, S. (2009). Cognition and renewal: Comparing ceo and organizational effects on incumbent adaptation to technical change. *Organization Science*, 20(2):461–477.
- Eklund, J. C. and Mannor, M. J. (2020). Keep your eye on the ball or on the field? exploring the performance implications of executive strategic attention. *Academy of Management Journal*, (ja).
- Eling, M. and Lehmann, M. (2018). The impact of digitalization on the insurance value chain and the insurability of risks. *The Geneva papers on risk and insurance-issues and practice*, 43(3):359–396.
- Eling, M. and Schmeiser, H. (2010). Insurance and the credit crisis: Impact and ten consequences for risk management and supervision. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 35(1):9–34.
- Fama, E. F. and Jensen, M. C. (1983). Separation of ownership and control. *The journal of law and Economics*, 26(2):301–325.
- Faraj, S., Jarvenpaa, S. L., and Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization science*, 22(5):1224–1239.
- Fisher, P. (2018). All companies are technology companies now.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.
- Flammer, C. and Bansal, P. (2017). Does a long-term orientation create value? evidence from a regression discontinuity. *Strategic Management Journal*, 38(9):1827–1847.
- Foote, C. L., Gerardi, K. S., and Willen, P. S. (2012). Why did so many people make so many ex post bad decisions? the causes of the foreclosure crisis. Technical report, National Bureau of Economic Research.
- Forman, C. and van Zeebroeck, N. (2019). Digital technology adoption and knowledge flows within firms: Can the internet overcome geographic and technological distance? *Research Policy*, 48(8):103697.
- Forum, W. E. (2016). The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution.
- Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.
- Gaba, V. and Joseph, J. (2013). Corporate structure and performance feedback: Aspirations and adaptation in m-form firms. *Organization Science*, 24(4):1102–1119.

- Gabel, S., Guhl, D., and Klapper, D. (2019). P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research*, 56(4):557–580.
- Gallus, J. and Bhatia, S. (2020). Gender, power and emotions in the collaborative production of knowledge: A large-scale analysis of wikipedia editor conversations. *Organizational Behavior and Human Decision Processes*, 160:115–130.
- Gątkowski, M., Dietl, M., Skrok, Ł., Whalen, R., and Rockett, K. (2020). Semantically-based patent thicket identification. *Research Policy*, 49(2):103925.
- Genetay, N. and Molyneux, P. (2016). *Bancassurance*. Springer.
- Geneva (2012). Insurance regulation — reflections for a post-crisis world.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
- Gibson, J. W. and Tesone, D. V. (2001). Management fads: Emergence, evolution, and implications for managers. *Academy of Management Perspectives*, 15(4):122–133.
- Glasserman, P. and Mamaysky, H. (2019). Does unusual news forecast market stress? *Journal of Financial and Quantitative Analysis*, 54(5):1937–1974.
- Goldberg, D. M., Khan, S., Zaman, N., Gruss, R. J., and Abrahams, A. S. (2020). Text mining approaches for postmarket food safety surveillance using online media. *Risk Analysis*.
- Gomaa, W. H., Fahmy, A. A., et al. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18.
- Goodman, L. A. (1960). On the exact variance of products. *Journal of the American statistical association*, 55(292):708–713.
- Gorbatai, A., Dioun, C., and Lashley, K. (2021). Making space for emotions: Empathy, contagion, and legitimacy's double-edged sword. *Organization Science*, 32(1):42–63.
- Grant, R. M. (2021). *Contemporary strategy analysis*. John Wiley & Sons.
- Grover, V. and Kohli, R. (2012). Cocreating it value: New capabilities and metrics for multifirm environments. *MIS Quarterly*, pages 225–232.
- Gründl, H., Post, T., and Schulze, R. N. (2006). To hedge or not to hedge: Managing demographic risk in life insurance companies. *Journal of Risk and Insurance*, 73(1):19–41.
- Guo, W., Sengul, M., and Yu, T. (2020). The impact of executive verbal communication on the convergence of investors' opinions. *Academy of Management Journal*, (ja).
- Guo, W., Yu, T., and Gimeno, J. (2017). Language and competition: Communication vagueness, interpretation difficulties, and market entry. *Academy of Management Journal*, 60(6):2073–2098.
- Hafez, P. A. (2011). How news events impact market sentiment. *The handbook of news analytics in finance*, page 143.

- Haldane, A. G. and May, R. M. (2011). Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355.
- Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., and Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13(2):586–632.
- HCLTech (2020). Thriving in a world of digital disruption.
- Heavey, C., Simsek, Z., Kyprianou, C., and Risius, M. (2020). How do strategic leaders engage with social media? a theoretical framework for research and practice. *Strategic Management Journal*, 41(8):1490–1527.
- Higham, K., De Rassenfosse, G., and Jaffe, A. B. (2021). Patent quality: towards a systematic framework for analysis and measurement. *Research Policy*, 50(4):104215.
- Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.
- Hoffman, A. J. and Ocasio, W. (2001). Not all events are attended equally: Toward a middle-range theory of industry attention to external events. *Organization science*, 12(4):414–434.
- Hopkin, P. (2018). *Fundamentals of risk management: understanding, evaluating and implementing effective risk management*. Kogan Page Publishers.
- Howard, A. (2010). Groupthink and corporate governance reform: Changing the formal and informal decisionmaking processes of corporate boards. *S. Cal. Interdisc. LJ*, 20:425.
- Huang, X., Zhou, H., and Zhu, H. (2009). A framework for assessing the systemic risk of major financial institutions. *Journal of Banking & Finance*, 33(11):2036–2049.
- Hund, A. (2020). Recombination in times of pervasive digitalization: a review.
- Hunjra, A. I., Hanif, M., Mehmood, R., and Nguyen, L. V. (2020). Diversification, corporate governance, regulation and bank risk-taking. *Journal of Financial Reporting and Accounting*.
- IAIS (2011). Insurance and financial stability.
- IASB (2007a). Preliminary views on insurance contracts, part 1: Invitation to comment and main text.
- IASB (2007b). Preliminary views on insurance contracts, part 2: Appendices.
- IFRS (2005). IFRS 4 insurance contracts.
- IFRS (2022). IFRS 17 insurance contracts.
- IMF, BIS, and FSB (2009). Guidance to assess the systemic importance of financial institutions, markets and instruments: Initial considerations.
- Ishwaran, H. (2015). The effect of splitting on random forests. *Machine learning*, 99(1):75–118.

- Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.
- Jackson, H. E. (1999). Regulation in a multisector financial services industry: An exploratory essay. *Wash. ULQ*, 77:319.
- Jacobides, M. G., Cennamo, C., and Gawer, A. (2018). Towards a theory of ecosystems. *Strategic management journal*, 39(8):2255–2276.
- Janis, I. L. (2008). Groupthink. *IEEE Engineering Management Review*, 36(1):36.
- Jarzabkowski, P., Bednarek, R., and Spee, P. (2015). *Making a market for acts of God: The practice of risk-trading in the global reinsurance industry*. Oxford University Press, USA.
- Jedidi, K., Schmitt, B. H., Sliman, M. B., and Li, Y. (2021). Express: R2m index 1.0: Assessing the practical relevance of academic marketing articles. *Journal of Marketing*, page 00222429211028145.
- Joseph, J. and Gaba, V. (2020). Organizational structure, information processing, and decision-making: a retrospective and road map for research. *Academy of Management Annals*, 14(1):267–302.
- Joseph, J. and Ocasio, W. (2012). Architecture, attention, and adaptation in the multibusiness firm: General electric from 1951 to 2001. *Strategic Management Journal*, 33(6):633–660.
- Joseph, J. and Wilson, A. J. (2018). The growth of the firm: An attention-based view. *Strategic Management Journal*, 39(6):1779–1800.
- Jurafsky, D. and Martin, J. H. (2021). Stanford University.
- Kao, H.-Y., Lin, S.-H., Ho, J.-M., and Chen, M.-S. (2004). Mining web informative structures and contents based on entropy analysis. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):41–55.
- Kaufman, G. G. (2014). Too big to fail in banking: What does it mean? *Journal of Financial Stability*, 13:214–223.
- Kellogg, K. C., Valentine, M. A., and Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1):366–410.
- Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20.
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- Kim, N., Lee, H., Kim, W., Lee, H., and Suh, J. H. (2015). Dynamic patterns of industry convergence: Evidence from a large amount of unstructured data. *Research Policy*, 44(9):1734–1748.
- Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.
- Klepper, S. (1997). Industry life cycles. *Industrial and corporate change*, 6(1):145–182.

- Ko, C.-W., Lee, J., and Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691.
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., and Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, 21(3):733–765.
- Kostovetsky, L. and Warner, J. B. (2020). Measuring innovation and product differentiation: Evidence from mutual funds. *The Journal of Finance*, 75(2):779–823.
- KPMG (2020). Kpmg connected enterprise.
- Krzeszewska, U., Poniszewska-Marańda, A., and Ochelska-Mierzejewska, J. (2022). Systematic comparison of vectorization methods in classification context. *Applied Sciences*, 12(10):5119.
- Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.
- Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.
- Kwon, O., An, Y., Kim, M., and Lee, C. (2020). Anticipating technology-driven industry convergence: evidence from large-scale patent analysis. *Technology Analysis & Strategic Management*, 32(4):363–378.
- Lang, M. and Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3):110–135.
- Lanzolla, G. and Markides, C. (2021). A business model view of strategy. *Journal of Management Studies*, 58(2):540–553.
- Lanzolla, G., Pesce, D., and Tucci, C. L. (2021). The digital transformation of search and recombination in the innovation function: Tensions and an integrative framework. *Journal of Product Innovation Management*, 38(1):90–113.
- Laureiro-Martinez, D. (2014). Cognitive control capabilities, routinization propensity, and decision-making performance. *Organization Science*, 25(4):1111–1133.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.
- Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising content and consumer engagement on social media: Evidence from facebook. *Management Science*, 64(11):5105–5131.
- Lee, I. and Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business horizons*, 61(1):35–46.
- Lee, J. and Berente, N. (2012). Digital innovation and the division of innovative labor: Digital controls in the automotive industry. *Organization Science*, 23(5):1428–1447.

- Lee, T. Y. and Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894.
- Leelaprute, P. and Amasaki, S. (2022). A comparative study on vectorization methods for non-functional requirements classification. *Information and Software Technology*, 150:106991.
- Leonardi, P. M. and Bailey, D. E. (2008). Transformational technologies and the creation of new work practices: Making implicit knowledge explicit in task-based offshoring. *MIS quarterly*, pages 411–436.
- Leonardi, P. M. and Barley, S. R. (2010). What’s under construction here? social action, materiality, and power in constructivist studies of technology and organizing. *Academy of Management Annals*, 4(1):1–51.
- Leonardi, P. M. and Vaast, E. (2017). Social media and their affordances for organizing: A review and agenda for research. *Academy of Management Annals*, 11(1):150–188.
- Li, Q., Maggitti, P. G., Smith, K. G., Tesluk, P. E., and Katila, R. (2013). Top management attention to innovation: The role of search selection and intensity in new product introductions. *Academy of Management Journal*, 56(3):893–916.
- Liu, J. and Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952.
- Liu, X., Wang, G. A., Fan, W., and Zhang, Z. (2020). Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis. *Information Systems Research*, 31(3):731–752.
- Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.
- Lundberg, S. M. and Lee, S.-I. (2017a). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.
- Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.
- Lusch, R. F. and Nambisan, S. (2015). Service innovation. *MIS quarterly*, 39(1):155–176.
- Lyytinen, K., Yoo, Y., and Boland Jr, R. J. (2016). Digital product innovation within four classes of innovation networks. *Information Systems Journal*, 26(1):47–75.
- Mabey, C. and Zhao, S. (2017). Managing five paradoxes of knowledge exchange in networked organizations: new priorities for hrm? *Human Resource Management Journal*, 27(1):39–57.
- Maiti, M. and Ghosh, U. (2021). Next generation internet of things in fintech ecosystem. *IEEE Internet of Things Journal*.
- Manning, C., Raghavan, P., and Schütze, H. (2008). Term weighting, and the vector space model. *Introduction to information retrieval*, pages 109–133.

- March, J. G., Olsen, J. P., and Christensen, S. (1976). *Ambiguity and choice in organizations*. Universitetsforlaget.
- March, JG, S. H. (1958). *Organizations*. Wiley, New York.
- Margalit, D., Rabinoff, J., and Rolen, L. (2017). Interactive linear algebra. *Georgia Institute of Technology*.
- Massa, L., Tucci, C., and Afuah, A. (2017). A critical assessment of business model research. *Academy of Management Annals*, 11(1):73–104.
- Masters, B. and Thiel, P. (2014). *Zero to one: notes on start ups, or how to build the future*. Random House.
- Maula, M. V., Keil, T., and Zahra, S. A. (2013). Top management's attention to discontinuous technological change: Corporate venture capital as an alert mechanism. *Organization Science*, 24(3):926–947.
- McKenny, A. F., Aguinis, H., Short, J. C., and Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management*, 44(7):2909–2933.
- McKinsey (2020). The state of ai in 2020. mckinsey global survey.
- Merz, M., Richman, R., Tsanakas, A., and Wuthrich, M. V. (2021). Interpreting deep learning models with marginal attribution by conditioning on quantiles. *Available at SSRN 3809674*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Milos, D., Sevic, Z., and Tekavčić, M. (2008). A review of the rationales for corporate risk management: Fashion or the need? *International Journal of Economic Sciences and Applied Research*, 1(1).
- Mishkin, F. S. (2007). *The economics of money, banking, and financial markets*. Pearson education.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
- Mousavi, R., Raghu, T., and Frey, K. (2020). Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums. *Journal of Management Information Systems*, 37(4):1073–1098.
- Nadkarni, S. and Barr, P. S. (2008). Environmental context, managerial cognition, and strategic action: An integrated view. *Strategic management journal*, 29(13):1395–1427.

- Nadkarni, S. and Chen, J. (2014). Bridging yesterday, today, and tomorrow: Ceo temporal focus, environmental dynamism, and rate of new product introduction. *Academy of Management Journal*, 57(6):1810–1833.
- NAIC (2011). Insurance and financial stability.
- NAIC (2013a). Buyer's guide for fixed deferred annuities.
- NAIC (2013b). Capital markets special report: Update on insurance industry investment portfolio asset mixes.
- NAIC (2013c). Capital markets special report: U.s. insurance industry's investment exposure to the financial sector.
- Nam, H., Joshi, Y. V., and Kannan, P. (2017). Harvesting brand information from social tags. *Journal of Marketing*, 81(4):88–108.
- Nambisan, S., Lyytinen, K., Majchrzak, A., and Song, M. (2017). Digital innovation management: Reinventing innovation management research in a digital world. *MIS Quarterly*, 41(1).
- Neale, F. R. and Drake, P. P. (2009). Financial guarantee insurance: Arrogance or ignorance in an era of exuberance.
- Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543.
- Newell, S., Scarbrough, H., and Swan, J. (2001). From global knowledge management to internal electronic fences: Contradictory outcomes of intranet development. *British Journal of Management*, 12(2):97–111.
- Nigam, A. and Ocasio, W. (2010). Event attention, environmental sensemaking, and change in institutional logics: An inductive analysis of the effects of public attention to clinton's health care reform initiative. *Organization Science*, 21(4):823–841.
- Ocasio, W. (1997). Towards an attention-based view of the firm. *Strategic Management Journal*, 18(S1):187–206.
- Ocasio, W. (2011). Attention to attention. *Organization Science*, 22(5):1286–1296.
- Ocasio, W. and Joseph, J. (2005). An attention-based theory of strategy formulation: Linking micro- and macroperspectives in strategy processes. In *Strategy Process*. Emerald Group Publishing Limited.
- Ocasio, W. and Radoynovska, N. (2016). Strategy and commitments to institutional logics: Organizational heterogeneity in business models and governance. *Strategic Organization*, 14(4):287–309.
- O'Connor, M. A. (2002). The enron board: The perils of groupthink. *U. Cin. L. Rev.*, 71:1233.
- Orlikowski, W. J. and Scott, S. V. (2008). 10 sociomateriality: challenging the separation of technology, work and organization. *Academy of Management Annals*, 2(1):433–474.

- Pan, Y., Huang, P., and Gopal, A. (2019). Storm clouds on the horizon? new entry threats and r&d investments in the us it industry. *Information Systems Research*, 30(2):540–562.
- Parker, G. G., Van Alstyne, M. W., and Choudary, S. P. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you*. WW Norton & Company.
- Patel, K. and Bhattacharyya, P. (2017). Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 31–36, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011a). Logistic regression in sklearn.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011b). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pentland, B. T. and Feldman, M. S. (2007). Narrative networks: Patterns of technology and organization. *Organization science*, 18(5):781–795.
- Pew (2020). Internet/broadband fact sheet.
- Philippe, D. and Durand, R. (2011). The impact of norm-conforming behaviors on firm reputation. *Strategic Management Journal*, 32(9):969–993.
- Piezunka, H. and Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations’ filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58(3):856–880.
- Piezunka, H. and Dahlander, L. (2019). Idea rejected, tie formed: Organizations’ feedback on crowdsourced ideas. *Academy of Management Journal*, 62(2):503–530.
- Piškorec, M., Antulov-Fantulin, N., Novak, P. K., Mozetič, I., Grčar, M., Vodenska, I., and Šmuc, T. (2014). News cohesiveness: an indicator of systemic risk in financial markets. *arXiv preprint arXiv:1402.3483*.
- Porter, M. E. (1985). Technology and competitive advantage. *Journal of business strategy*.
- Porter, M. E. and Heppelmann, J. E. (2015). How smart, connected products are transforming companies. *Harvard business review*, 93(10):96–114.
- Porter, M. E. P. M. (1986). *Competition in global industries*. Harvard Business Press.

- Powell, W. W., Horvath, A., and Brandtner, C. (2016). Click and mortar: Organizations on the web. *Research in Organizational Behavior*, 36:101–120.
- Power, M. (2009). The risk management of nothing. *Accounting, organizations and society*, 34(6-7):849–855.
- Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Preacher, K. J. and Hayes, A. F. (2004). Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior research methods, instruments, & computers*, 36(4):717–731.
- Puranam, P., Stieglitz, N., Osman, M., and Pillutla, M. M. (2015). Modelling bounded rationality in organizations: Progress and prospects. *Academy of Management Annals*, 9(1):337–392.
- PwC (2020a). Accelerate digital transformation.
- PwC (2020b). The digital insurance workforce: Going beyond aspirations.
- Rajaraman, A. and Ullman, J. D. (2011). Data mining (pp. 1–17).
- Reinhart, C. M. and Rogoff, K. S. (2009). This time is different. In *This Time Is Different*. princeton university press.
- Rerup, C. (2009). Attentional triangulation: Learning from unexpected rare crises. *Organization Science*, 20(5):876–893.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Rubio-Misas, M. (2022). Bancassurance and the coexistence of multiple insurance distribution channels. *International Journal of Bank Marketing*.
- Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., and Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, 104(10):1207.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.
- Schich, S. (2008). Challenges related to financial guarantee insurance. *OECD Journal: Financial Market Trends*, 2008(1):81–113.
- Schilling, M. A. (2000). Toward a general modular systems theory and its application to interfirm product modularity. *Academy of Management Review*, 25(2):312–334.
- Schmiedel, T., Müller, O., and Vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, 22(4):941–968.

- SEC (2020). U.s. credit markets: Interconnectedness and the effects of the covid-19 economic shock.
- SEC (2021). How to read a 10-k/10-q, u.s. securities and exchange commission.
- SEC, U. S. and Commission, E. (2011). How to read a 10-k.
- Sehlstedt, U., Bohlin, N., de Maré, F., and Beetz, R. (2016). Embracing digital health in the pharmaceutical industry. *International Journal of Healthcare Management*, 9(3):145–148.
- Shapley, L. (1953). A value for n -person games. In *Contributions to the Theory of Games*, pages 307–317. Princeton: Princeton University Press.
- Shepherd, D. A., McMullen, J. S., and Ocasio, W. (2017). Is that an opportunity? an attention model of top managers' opportunity beliefs for strategic action. *Strategic Management Journal*, 38(3):626–644.
- Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods*, 7(4):422.
- Sigma (2007). Bancassurance: Emerging trends, opportunities and challenges, sigma no. 5/2007.
- Simon, H. A. (1979). Rational decision making in business organizations. *The American economic review*, 69(4):493–513.
- Simpson, R. E. (1974). *Introductory electronics for scientists and engineers*. Allyn & Bacon.
- Singh, A. K. and Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10(7).
- Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.
- Sjostrom, Jr, W. K. (2009). The aig bailout. *Washington and Lee Law Review*, 66:943.
- Song, R., Moon, S., Chen, H. A., and Houston, M. B. (2018). When marketing strategy meets culture: the role of culture in product evaluations. *Journal of the Academy of Marketing Science*, 46(3):384–402.
- Stanford (2009a). Stemming and lemmatization.
- Stanford (2009b). Stemming and lemmatization, stanford nlp group.
- Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of management review*, 20(3):571–610.
- Sullivan, B. N. (2010). Competition and beyond: Problems and attention allocation in the organizational rulemaking process. *Organization Science*, 21(2):432–450.

- Taeuscher, K., Bouncken, R., and Pesch, R. (2021). Gaining legitimacy by being different: Optimal distinctiveness in crowdfunding platforms. *Academy of Management Journal*, 64(1):149–179.
- Tarashev, N., Tsatsaronis, K., and Borio, C. (2016). Risk attribution using the shapley value: Methodology and policy applications. *Review of Finance*, 20(3):1189–1213.
- Teece, D. J. (2010). Business models, business strategy and innovation. *Long Range Planning*, 43(2-3):172–194.
- Teece, D. J. (2018). Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world. *Research Policy*, 47(8):1367–1387.
- Theurer, C. P., Schäpers, P., Tumasjan, A., Welp, I., and Lievens, F. (2021). What you see is what you get? measuring companies' projected employer image attributes via companies' employment webpages. *Human Resource Management*.
- Thimann, C. (2015). Systemic features of insurance and banking, and the role of leverage, capital and loss absorption. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 40(3):359–384.
- Timoshenko, A. and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1):1–20.
- Tobias, A. and Brunnermeier, M. K. (2016). Covar. *The American Economic Review*, 106(7):1705.
- Toubia, O. and Netzer, O. (2017). Idea generation, creativity, and prototypicality. *Marketing science*, 36(1):1–20.
- Tripsas, M. and Gavetti, G. (2000). Capabilities, cognition, and inertia: Evidence from digital imaging. *Strategic Management Journal*, 21(10-11):1147–1161.
- Trueman, B. (1994). Analyst forecasts and herding behavior. *The review of financial studies*, 7(1):97–124.
- Tsanakas, A. and Millossovich, P. (2016). Sensitivity analysis using risk measures. *Risk Analysis*, 36(1):30–48.
- Tuggle, C. S., Sirmon, D. G., Reutzel, C. R., and Bierman, L. (2010). Commanding board of director attention: investigating how organizational performance and ceo duality affect board members' attention to monitoring. *Strategic management journal*, 31(9):946–968.
- Turber, S., Brocke vom, J., and Gassmann, O. (2015). Designing business models in the age of pervasive digitization. 2015(1):11600.
- Turk, M. C. (2015). The convergence of insurance with banking and securities industries, and the limits of regulatory arbitrage in finance. *Columbia Business Law Review*, page 967.
- Tziogkidis, P., Philippas, D., and Tsionas, M. G. (2020). Multidirectional conditional convergence in european banking. *Journal of Economic Behavior & Organization*, 173:88–106.

- Uzunca, B. (2018). A competence-based view of industry evolution: The impact of sub-market convergence on incumbent- entrant dynamics. *Academy of Management Journal*, 61(2):738–768.
- Vanacker, T., Forbes, D. P., Knockaert, M., and Manigart, S. (2020). Signal strength, media attention, and resource mobilization: evidence from new private equity firms. *Academy of Management Journal*, 63(4):1082–1105.
- von Hippel, E. and Kaulartz, S. (2021). Next-generation consumer innovation search: Identifying early-stage need-solution pairs on the web. *Research Policy*, 50(8):104056.
- Vuori, T. O. and Huy, Q. N. (2016). Distributed attention and shared emotions in the innovation process: How nokia lost the smartphone battle. *Administrative Science Quarterly*, 61(1):9–51.
- Wallison, P. J. (2008). Convergence in financial services markets: likely effects on insurance regulation. *American Enterprise Institute*.
- Wang, Q., Li, B., and Singh, P. V. (2018). Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research*, 29(2):273–291.
- Wang, X., He, J., Curry, D. J., and Ryoo, J. H. (2021). Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews. *Journal of Marketing*, page 00222429211047822.
- Weick, K. E. and Sutcliffe, K. M. (2006). Mindfulness and the quality of organizational attention. *Organization science*, 17(4):514–524.
- Weill, L. (2009). Convergence in banking efficiency across european countries. *Journal of International Financial Markets, Institutions and Money*, 19(5):818–833.
- Yoo, Y., Boland Jr, R. J., Lyytinen, K., and Majchrzak, A. (2012). Organizing for innovation in the digitized world. *Organization Science*, 23(5):1398–1408.
- Zhao, R. and Mao, K. (2017). Fuzzy bag-of-words model for document representation. *IEEE transactions on fuzzy systems*, 26(2):794–804.
- Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., and Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems*, 35(2):510–539.
- Zott, C. and Amit, R. (2007). Business model design and the performance of entrepreneurial firms. *Organization science*, 18(2):181–199.

Appendix A

Constructing Blurring

A more formal definition of *Blurring* is given here. Consider a set of firms, each of which can belong to one of M categories, $\{C_1, \dots, C_M\}$. Denote the text data of firm i in year t by D_{it} . Consider firm i , which belongs to category C_k , i.e. $i \in C_k$. A classification algorithm is used to estimate the probability that the report D_{it} belongs to each of the M categories; denote these estimated probabilities as $\hat{P}(i \in C_j | D_{it})$, for $j=1, \dots, M$. Subsequently the *Blurring* for firm i in year t is defined as:

$$Blur_{it} = 1 - \hat{P}(i \in C_k | D_{it}). \quad (A.1)$$

This definition easily accommodates the situation where there are more than two types: then *Blurring* as defined in A.1 equals the sum of the predicted probabilities of the wrong types, that is, $Blur_{it} = \sum_{j \neq k} \hat{P}(i \in C_j | D_{it})$.

The text data D_{it} are represented as follows. The vocabulary is the list of all terms in our dataset. To avoid endogeneity, any terms related to digitalization, which are used to construct the independent variable are excluded. Because of the low frequency of digitalization words, removing these does not substantially alter the rest of the text. D_{it} then is a vector, of the same length as the vocabulary, each element of which represents the relative frequency of a given term occurring in the report of firm i in year t . Hence the first element of D_{it} equals the number of times the first term in the vocabulary appears in the report of firm i in year t , divided by the total number of terms in that report (hence the elements of D_{it} sum to one). As the length of the documents varies, the normalization ensures that classification operates on the composition of documents' vocabulary rather than their length.

Before training the classification algorithm, a final processing step needs to be carried out on the data: rebalancing. Our data sample is very unbalanced, with most observations dominated by banks, especially Credit Banks. Unbalanced data can lead to suboptimal

classification performance (Chawla et al., 2004). Consequently, we balance the samples using Synthetic Minority Over-Sampling Technique (SMOTE), involving resampling from all classes but the majority class (Chawla et al., 2002).

Blurring is then calculated by the following process:

1. Remove all the digitalization and business models keywords, listed in Appendix B, from the dataset document D_{it} .
2. Estimate a regularized logistic regression model (Pedregosa et al., 2011b) using all remaining documents in year t .
3. Use the estimated model to calculate the predicted probabilities $\hat{P}(i \in C_j | D_{it})$.
4. Work out $Blur_{it}$ from equation A.1.

The first step is introduced to ensure that predictions on the class of D_{it} are ‘out-of-sample’. Regularization of the logistic regression model addresses the large size of the vocabulary (in our data, we have 43798 items in the vocabulary) by penalizing over-parameterization.

We use Python for the above process of constructing the dependent variable. The text data are pre-processed using the Python packages spaCy and NLTK. Then they are vectorized and rebalanced using sklearn and imblearn respectively. Finally, the logistic regression and the calculation of *Blurring* are done using sklearn. We use the default regularization parameter of $C = 1$ for all reported results; in robustness checks we consider a higher degree of regularization ($C = 0.01$) (Pedregosa et al., 2011a).

Appendix B

Full lists of keywords of digitalization and business models

The full extended list of the 61 keywords of *Digitalization* are as follows:

online, web, digits, information, networking, websites, video, cyberspace, analog, broadband, press, networks, audio, software, wireless, database, phone, aol, algebraic, google, computer, analysis, sugiyama, facebook, computers, network, messaging, analytical, electronic, statistics, cameras, microsoft, multimedia, website, security, youtube, technology, yahoo, sites, download, phones, mail, cable, blogs, double-digit, robotics, interactive, tracking, camera, mobile, databases, chat, computerization, browser, empirical, media, internet, subscribers, messages, formats, downloads.

The full extended list of the 276 keywords of *Business Models* are as follows:

revenues, costs, offered, functions, networks, profits, models, offering, offer, structures, ecosystems, create, markets, organisation, designs, offers, prices, chains, incentives, ventures, policies, services, relationships, systems, segments, products, benefits, funds, opportunity, delivering, components, operations, earnings, resources, loops, customers, capturing, bilateral, architectural, delivered, captured, economy, supplier, organizations, units, partners, cash, strategies, created, costing, pretax, socio-economic, advertising, performances, businesses, efficacy, activities, creates, connections, manufacturers, link, relations, trading, net, delivers, values, processes, stock, creation, creating, designed, environments, growth, regulations, industry, socio, efficient, distributors, production, deliveries, personnel, financial, economic, exchanges, stores, investors, design, architecture, revenue, profit, pay, expensive, designing, cooperate, income, fund, offerings, reasoning, strategy, policy, departments, environmental, economies, benefiting, sales, regulatory, nonprofit, plan, links, input, price, cost, crisis, gains, regulate, transparency, efficiencies, ties, stocks, paying, valued, functional, regulating,

coordination, selling, promote, channel, possibilities, importers, employees, subcontractors, designers, paid, agreements, development, market, administration, marketplace, partnership, unprofitable, investment, funding, accountability, non-profit, connected, buying, benefited, costly, tactics, distribute, organisations, division, agreement, business, change, decisions, internal, external, employee, complex, co-operation, operating, product, marketing, institutions, value, monetary, dealers, expense, distributed, regulated, financing, invest, proposal, rewards, retailers, traders, negotiations, tax, structural, productivity, beneficiaries, buyers, internet, sustainability, credit, strategic, cooperation, innovative, commodity, captures, reliability, agenda, establishment, taxpayer, promotional, recapture, global, pricing, reward, relation, lucrative, forecast, consumption, distributing, stability, producers, competitiveness, consumers, effort, vendors, prosperity, consumer, bonuses, downturn, subsidiary, architects, premium, payments, constellations, outputs, assets, profitable, profitability, channels, performing, collaboration, gain, sustainable, governmental, stakeholders, promoting, invested, priced, manufacturing, propositions, quality, demand, sell, legislation, efficiency, effectiveness, substantial, commercial, communications, modeling, cooperative, functioning, organized, contributions, export, brands, investments, participation, exporting, model, emerging, chance, effective, negotiation, laws, prototype, enterprise, interrelation, viability, integration, organizational, interdependence, technology, distributions, imports, interaction, licensing, asset, interrelatedness, contractors.

Appendix C

Literature review of text similarity in management research

Below are two tables summarizing all the research reviewed in Section 2.2. Table C.1 summarizes all the research using text similarity to directly measure the management constructs themselves, while Table C.2 shows those using text similarity to measure the cross-sectional or temporal similarity or difference of the same management constructs. The difference between the two types are discussed at the beginning of Section 2.2.1. The benchmarks of all the examples in Table C.2 are the documents of other individuals or periods. Due to the simplicity, the benchmarks are not shown in Table C.2.

Table C.1 Examples of Management Constructs Directly Measured with Text Similarity.

Management constructs	Data source	Benchmarks	Measures of similarity	Reference
Innovation, novelty	Web crawler collected the text content of patents from the USPTO patent search website	Past patents	Backward inverse document frequency of the keywords of technologies	Kelly et al. (2021)
	Patent titles, abstracts and claims for all granted U.S. utility patents from the USPTO, the patent claims research dataset, and PATSTAT	Past patents	Cosine similarity	Arts et al. (2021)
	English posts associated to the kiteboarding domain from over 9617 different websites across the globe	A list of innovation concepts, including various spelling variations of each aspect, and semantically similar words suggested by word-space model	Cosine similarity	von Hippel and Kaulartz (2021)
Continued on next page				

Table C.1 – continued from previous page

Management constructs	Data source	Benchmarks	Measures of similarity	Reference
Customer agility	The cohort of patents granted by the USPTO	All the patents are benchmarks of each other	Cosine similarity	Higham et al. (2021)
	Posts and comments submitted by all users on Reddit	The same user's own past awarded content	Cosine similarity	Burtch et al. (2021)
	The top 50 search results on Google using the exact wording of the idea generation topic as the text of the query	All the word stems appeared in the search results are benchmarks of each other, forming a network	Jaccard indexes	Toubia and Netzer (2017)
	Apple App Store's daily product data, and online review data from customers accordingly	The release note of a sequential innovation	Cosines similarity	Zhou et al. (2018)
Quality of idea	Q&As in Yahoo! Answers health section	Answers provided by health care professionals	Cosine similarity	Mousavi et al. (2020)
Continued on next page				

Table C.1 – continued from previous page

Management constructs	Data source	Benchmarks	Measures of similarity	Reference
Intellectual distance / experience relevance	A list of potential employers in the German laser industry, a complete list of laser-related patent applications by these firms, the complete record of doctoral dissertations defended at German universities, as well as scientific publications related to these dissertations	Patent portfolios of the firms or the inventors	Cosine similarity	Buenstorf and Heinisch (2020)
	Application information of external applicants for teaching positions at the Minneapolis Public School District	Job description	Euclidean (L2) distance	Sajjadiani et al. (2019)
Continued on next page				

Table C.1 – continued from previous page

Management constructs	Data source	Benchmarks	Measures of similarity	Reference
Female-typedness	Monsterboard and Textkernel, which provided access to vacancy data from various employment websites	Job description	Cosine similarity	Kobayashi et al. (2018)
	Expertise descriptions of all of the organization's employees, and R&D project applications in the firm's R&D project management system	574 keywords on expertise, identified ex ante	The number of overlapping keywords between the selector's skill description and the R&D project proposal	Criscuolo et al. (2017)
	Wikipedia Talk page discussions, including comments and articles	20 male words relative to a set of 20 female words, e.g., gender pronouns, gendered relationship words	Cosine similarity	Gallus and Bhatia (2020)
Continued on next page				

Table C.1 – continued from previous page

Management constructs	Data source	Benchmarks	Measures of similarity	Reference
Market structure	Consumer messages on sedan cars from the Sedans Forum on Edmunds.com	Consumer messages are benchmarks of each other	Lift or pointwise mutual information, Jaccard index, Cosines similarity, TFIDF, Pearson correlation	Netzer et al. (2012)
	Reviews for all digital cameras available at Epinions.com	All the phrases. Phrases are automatically grouped together according to their similarity so as to elicit product attributes	Cosine similarity	Lee and Bradlow (2011)
Unusualness of news	Thomson Reuters news articles about the top 50 global banks, insurance, and real estate firms	All the other news	Entropy, the ratio of the number of words that two articles have in common to the number of distinct words occurring in the two articles combined	Glasserman and Mamaysky (2019)

Table C.2 Temporal and cross-sectional concepts measured by text similarity.

Management constructs	Measures of similarity	References
A. Cross-sectional similarity		
Distinctiveness between an entrepreneurial story deviates from the market category's prototypical entrepreneurial story while crowdfunding	Difference between the individual and industry average document vectors	Taeuscher et al. (2021)
Convergence / divergence of opinions /ideas, agreement / disagreement	Cosine similarity, absolute difference of each sentiment score Cosine similarity Kullback-Liebler divergence Pointwise mutual information Cosine similarity	Deng et al. (2021) Guo et al. (2020) Cookson and Niessner (2020) Archak et al. (2011) Wang et al. (2018)
Match / relevance of content between the initial idea and response	Cosine similarity, Kullback–Leibler divergence, entropy, and perplexity	Liu et al. (2020)
Relevance of academic research to marketing practice	The sum of LDA topic probability distribution, weighted by marketing-term relevance scores obtained from business practitioners Cosine similarity	Jedidi et al. (2021) Liu and Toubia (2018); Piezunka and Dahlander (2015, 2019)
Similarity of patent written description to claims	Cosine similarity	Ashtor (2022)
Continued on next page		

Table C.2 – continued from previous page

Management constructs	Measures of similarity	References
Relationship of a product's technical specifications and abstract product benefits	Cosine similarity	Wang et al. (2021)
Product uniqueness	Cosine similarity	Pan et al. (2019)
Product similarity	t-SNE	Gabel et al. (2019)
	Cosine similarity	Abbas et al. (2018); Hoberg and Phillips (2016)
Personality similarity	Ratio of personality language scores	Akstinaite et al. (2020)
	Jaccard similarity coefficient	Adamopoulos et al. (2018)
Cultural congruence	Begin with a list of cultural terms elicited from the movie reviews, and identify four predefined cultural categories	Song et al. (2018)
Culture heterogeneity	Jensen–Shannon divergence	Corritore et al. (2020)
Industry convergence	Normalized pointwise mutual information (PMI), co-occurrence	Kim et al. (2015)
Similarity of brands' social tags	Cosine similarity	Nam et al. (2017)
Coherence of collective identity	Cosine similarity	Gorbatai et al. (2021)
Bibliographic coupling network	Cosine similarity	Chen et al. (2019)
Comparability of user-generated-content-based customer needs and interview-based customer needs	If the clustering (CNN) worked perfectly, samples within each of the clusters would articulate the same customer need	Timoshenko and Hauser (2019)

Continued on next page

Table C.2 – continued from previous page

Management constructs	Measures of similarity	References
Firm co-benefit/counter-benefit network	Sentiment classification (positive or neagative). If two firms have the same (different) sentiment, they co-benefit (counter-benefit)	Chen et al. (2021)
Relationship between risk types	Silhouette value	Bao and Datta (2014)
Relationship of the firm and the underlying news story	RavenPack score, a third party score	Di and Laux (2021)
Distinctiveness of projected employer image attributes among companies / industries; Convergence with third-party employer ratings	Silhouette value, regression coefficients	Theurer et al. (2021)
B. Temporal similarity		
Changes in organizational identity over time	Jaccard similarity coefficient	Powell et al. (2016)
Consistency of employer image attributes over time	ANOVA	Theurer et al. (2021)
The effect of social media advertising content on customer engagement	The extent of serial correlation in post content, measured by XOR / Exclusive OR	Lee et al. (2018)

Appendix D

Definition of text similarity measures

In this appendix, we provide the definition of the text similarity measures introduced in Section 2.2.3.

D.1 Overlapping terms

We define $\mathbf{A} = \{f_{A1}, f_{A2}, \dots, f_{Ap}\} \in \mathbb{R}^{1 \times p}$ and $\mathbf{B} = \{f_{B1}, f_{B2}, \dots, f_{Bp}\} \in \mathbb{R}^{1 \times p}$ as two document vectors, where p is the number of features. p can be, for example, the size of the vocabulary for Term Frequency vectorization, or the number of topics for topic modelling vectorization.

Jaccard similarity is defined in Equation (D.3) and (D.4). In Equation (D.3), $|\mathbf{A} \cap \mathbf{B}|$ and $|\mathbf{A} \cup \mathbf{B}|$ represent the overlapping features and the total number of features respectively. If we focus on terms, $|\mathbf{A} \cap \mathbf{B}|$ is the count of overlapping terms in the two documents $\mathbf{A}$ and $\mathbf{B}$, while $|\mathbf{A} \cup \mathbf{B}|$ is the total size of the vocabulary of the two documents. We can also replace "terms" with "topics", and then $|\mathbf{A} \cap \mathbf{B}|$ is the total percentage of overlapping topics in the two documents, while $|\mathbf{A} \cup \mathbf{B}|$ is the total number of topics.

Equation (D.4) define Jaccard similarity more specifically. The count of overlapping features is defined as follows. We first define the overlapping vector,

$$OL(\mathbf{A}, \mathbf{B}) = \{ol_{AB1}, ol_{AB2}, \dots, ol_{ABp}\} \in \mathbb{R}^{1 \times p}, \quad (\text{D.1})$$

where

$$ol_{ABj} = \min\{f_{Aj}, f_{Bj}\}, \quad j \in \{1, \dots, p\}. \quad (\text{D.2})$$

Then the Jaccard Similarity of **A** and **B** is:

$$Jaccard\ Similarity(\mathbf{A}, \mathbf{B}) = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|} \quad (\text{D.3})$$

$$= \frac{\sum_{j=1}^p ol_{ABj}}{p}. \quad (\text{D.4})$$

Applying *Exclusive OR* or *XOR* to text similarity, the XOR vector is defined as:

$$\mathbf{A} \text{ XOR } \mathbf{B} = \{x_1, x_2, \dots, x_p\} \in \mathbb{R}^{1 \times p}, \quad (\text{D.5})$$

where

$$x_j = \begin{cases} 1 & \text{if } f_{Aj} \neq f_{Bj}, \\ 0 & \text{if } f_{Aj} = f_{Bj}, \end{cases} \quad (\text{D.6})$$

$j \in \{1, \dots, p\}$.

The text dissimilarity is defined as the mean of the XOR vector:

$$\text{The mean of the XOR vector} = \frac{1}{p} \sum_{j=1}^p x_j. \quad (\text{D.7})$$

D.2 Vector distance / angle

The *Euclidean distance* between the two document vectors are:

$$Euclidean\ distance = \sqrt{\sum_{j=1}^p (f_{Aj} - f_{Bj})^2}. \quad (\text{D.8})$$

Cosine similarity of the two documents is defined as:

$$Cosine\ Similarity = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_{j=1}^p f_{Aj} f_{Bj}}{\sqrt{\sum_{j=1}^p f_{Aj}^2} \sqrt{\sum_{j=1}^p f_{Bj}^2}}. \quad (\text{D.9})$$

D.3 Information-based measures

Now suppose we have N document vectors. Based on the definition in Kao et al. (2004), the *entropy* of the j th feature ($j \in \{1, \dots, p\}$) is:

$$e(j) = - \sum_{i=1}^N \tilde{f}_{ij} \log_N \tilde{f}_{ij}, \quad (\text{D.10})$$

in which $\tilde{f}_{ij}$ is the value of normalized value of the j th feature in the i th document, which is exactly the percentage of feature j (e.g., the percentage of term frequency), $i \in \{1, \dots, N\}$, i.e.

$$\tilde{f}_{ij} = \frac{f_{ij}}{\sum_{i=1}^N f_{ij}} \in [0, 1] \quad (\text{D.11})$$

Then the entropy of the i th document vector is the weighted summation of its features' entropies:

$$e(i) = \sum_{j=1}^p \tilde{f}_{ij} e(j). \quad (\text{D.12})$$

The definition in Equation (D.12) correct Kao et al. (2004)'s definition of entropy of a document by adding the weight $\tilde{f}_{ij}$.

Perplexity is a function of entropy. The perplexity of the j th feature is:

$$pp(j) = 2^{e(j)}. \quad (\text{D.13})$$

Then the perplexity of the i th document vector is:

$$pp(i) = \sum_{j=1}^p \tilde{f}_{ij} \cdot pp(j), \quad (\text{D.14})$$

where $e(j)$ is defined in Equation (D.10).

If we focus on two of the document vectors, $\mathbf{A}$, $\mathbf{B}$, i.e. $i = A$ or $B \in \{1, \dots, N\}$, then the *Kullback-Leibler Divergence* of the two vectors is defined as:

$$D_{KL}(\mathbf{A}||\mathbf{B}) = \sum_{j=1}^p \tilde{f}_{Aj} \cdot \log \left(\frac{\tilde{f}_{Aj}}{\tilde{f}_{Bj}} \right). \quad (\text{D.15})$$

Here, the feature distribution of $\mathbf{B}$ is taken as the reference, and the expectation is taken using the feature distribution of $\mathbf{A}$. We can also take the feature distribution of $\mathbf{A}$ as the reference, and the results will be different.

To overcome the asymmetry of Kullback-Leibler Divergence, *Jensen-Shannon Divergence* provides a symmetric measure defined as:

$$D_{JS}(\mathbf{A}||\mathbf{B}) = \frac{1}{2}D_{KL}(\mathbf{A}||\mathbf{M}) + \frac{1}{2}D_{KL}(\mathbf{B}||\mathbf{M}), \quad (\text{D.16})$$

where $\mathbf{M} = \frac{1}{2}(\mathbf{A} + \mathbf{B})$.

If we focus on two features, say, the u th and v th feature, $u, v \in \{1, \dots, p\}$ (the features are usually terms), then the *Pointwise Mutual Information* (PMI) or *lift* between the two features is defined as follows based on Netzer et al. (2012):

$$PMI(u, v) = \frac{P(u, v)}{P(u) \times P(v)}, \quad (\text{D.17})$$

where $P(u, v)$ is the percentage of vectors containing both the u th and v th features, and $P(u)$ or $P(v)$ is the percentage of vectors containing the u th or v th feature respectively. There can be different variations of the definition of *PMI* to measure text similarity. For example, Archak et al. (2011) use the count rather than percentage to calculate *PMI*.