Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 1 Capitole (UT1 Capitole)

Si vous êtes en cotutelle internationale, remplissez ce champs en notant : Cotutelle internationale avec " nom de l'établissement", sinon effacer ce texte pour qu'il n'apparaisse pas à l'impression

---

Présentée et soutenue par :
XIAOJUAN LIU
le jeudi 14 septembre 2023

Titre :
Essays in corporate finance based on textual analysis

École doctorale et discipline ou spécialité :

ED SG : Finance

Unité de recherche :
TSM research

Directeur/trice(s) de Thèse :
Madame,Catherine,Casamatta, Professeur de Université, Toulouse School of Economics
Monsieur,David,Stolin, Professeur de Université, Toulouse Business School

Jury :
Helen Bollaert, professor, SKEMA Business School (rapporteur)
Catherine Casamatta,professor, TSM, Université Toulouse Capitole
Sonia Jimenez-Garces, professor, Grenoble INP, Université Grenoble Alpes (rapporteur)
Sophie Moinas, professor, TSM, Université Toulouse Capitole
Konrad Raff, associate professor, Norwegian School of Economics
David Stolin,professor,Toulouse Business School

Acknowledgement

Embarking on the voyage to attain my Ph.D. has been a monumental undertaking, a journey teeming with challenges and growth opportunities that I could not have navigated without the aid of numerous influential individuals. As I reflect on this pivotal period in my life, I am awash with gratitude for the many individuals who have significantly influenced my journey.

Firstly, I wish to express profound appreciation to my advisors, Prof. Catherine Casamatta and Prof. David Stolin, whose mentorship was priceless. They perpetually challenged me to expand my perspective, delve deeper into my research, and accept nothing short of my best efforts. Their intellectual rigor, unwavering commitment, and constant encouragement have been pivotal to my personal and professional evolution.

Prof. Catherine Casamatta has been my guiding light since my Master's program. For the past seven years, her patience with me has been unwavering. Despite the trials and tribulations I encountered, she ceaselessly directed me towards the right path. In my scientific research journey, she has served as a constant role model. Prof. David Stolin, from our first encounter, has been my mentor in life, academics, and crucial decision-making, akin to a paternal figure. His steadfast presence and guidance, especially in moments of uncertainty, have been instrumental in my doctoral journey. To him, I want to express, "Look, I finally made it."

I also wish to acknowledge two significant individuals who have greatly impacted my life. Firstly, Prof. Gerard Hoberg, who extended an invitation for me to become a visiting scholar at the University of Southern California. His teachings in text analysis research methodologies, coupled with his affable personality and empathetic demeanor, made my time at USC enriching and enjoyable. Regardless of what lies ahead, I am genuinely thankful for his influence in my life. Secondly, Dr. Dakang Huang, who has been an invaluable friend during my Ph.D. journey and he got his Ph.D. at TSE this summer. His optimistic outlook towards academia and life has been truly inspirational. Our shared discussions on research and life planning have been enlightening, and his positive approach to problem-solving is something I deeply admire. Witnessing his successful completion of his Ph.D. has been a delight, and I am thrilled to be on the cusp of reaching the same milestone.

I would be remiss not to express my deep gratitude towards all the professors from whom I had the privilege to learn and collaborate at TSM and TSE. Their extensive knowledge and fervor for their respective fields ignited my inspiration and have had a lasting impact on my academic trajectory. I thank Milo Bianchi, Christophe Bisière, Sylvain Bourjade, Matthieu Bouvard, Patrick Coen, Laurent Fresard, Ulrich Hege,

Sophie Moinas, Silvia Rossetto.

My colleagues and friends have also played a key role in this journey, and for this, I offer my sincere thanks. The list includes but is not limited to Li Bao, Mudit Dhakar, Yue Fei, Diana Castro Herrera, Yixin Huang, Phuong Khanh Huynh, Natalia Kovaleva, Hien Pham, Yovin Sadasing, Keller Martinez Solis, Maxime Wavasseur, Jun Yan, Li Yu, Suxiu Yu, and Yifei Zhang. Our camaraderie, intellectual discourse, and unwavering support made this endeavor not only manageable but also pleasurable. Your friendship provided solace and resilience during tough times.

Finally, my deepest gratitude goes to my family. Your unconditional love, unwavering support, and continual encouragement have been the bedrock of my journey. Your faith in me, even during moments of self-doubt, is something I will forever cherish.

This acknowledgment only begins to touch on the depth of my gratitude. Each one of you has indelibly shaped my journey in a unique way, and for this, I extend my heartfelt thanks. Achieving this milestone would have been unattainable without your collective contributions.

Introduction

The main topic of my thesis which consists of three chapters is the utilization of textual data analysis to explore various aspects of finance and industrial organization. Specifically, Chapter 1 focuses on the development of a dynamic global text-based industry classification that surpasses traditional industry classifications in generating homogeneous groups of firms, particularly in emerging economies. Chapter 2 examines the role of product relatedness in mergers and acquisitions on the global market, utilizing textual analysis to construct measures of pairwise product similarity. Finally, Chapter 3 investigates the impact of news across countries and languages on stock market indices, employing a comprehensive dataset of news articles and multi-lingual text processing techniques to assess the value of news in explaining market volatility.

My work uses textual analysis, a relatively new tool in the finance and accounting fields, to investigate research questions in finance. This emerging field encompasses various branches, including document similarity (Hoberg and Phillips, 2016) and sentiment analysis, as well as readability (Lang and Lawrence, 2015). My research specifically contributes to the domain of document similarity. The fundamental idea behind the text-based approach is to utilize textual materials to construct variables, providing a fresh perspective on finance and economics (Gentzkow, Kelly, and Taddy, 2019; Ash and Hansen, 2023). While traditionally, researchers have relied on numerical figures to investigate their research questions, a wealth of valuable information is embedded in textual data sources. These sources include annual reports, news articles covering transactions, conference calls, and more. Textual analysis serves as a medium through which researchers can extract and interpret the rich information contained within texts. By employing advanced techniques, such as natural language processing and machine learning algorithms, textual analysis enables researchers to analyze and uncover insights from textual data. It allows for the exploration of firms' performance and the financial market by delving into the information concealed within textual sources. This approach expands the analytical toolkit beyond traditional numerical analysis and offers new avenues for understanding and interpreting financial phenomena.

An example of a prominent recent study using textual analysis is Bandiera et al. (2020) which investigate how CEOs' behavior can influence firm performance. The text data used by authors are daily phone call logs. They find that CEOs spend most of their time interacting with firms' insiders such as finance and marketing staff. They apply the latent Dirichlet allocation (LDA) method to build the behavior index of

CEOs and they uncover two different types which are leaders and managers. They find, on average, firms led by CEO leaders demonstrate higher levels of productivity, and this disparity becomes apparent only after the CEO assumes their position within the company. This is one illustration from a burgeoning literature that applies advanced text method to address a wide range of financial and economic questions (Gentzkow, Shapiro, and Taddy, 2019; Bertrand et al. 2021; Shapiro and Wilson, 2022)

My dissertation seeks to contribute to this literature and is organized in three chapters.

Chapter 1 provides an overview of the importance of accurately categorizing firms into industries for research in finance and industrial organization. The chapter highlights the limitations of standard industry classifications, such as the SIC code, in providing comprehensive information about the competitive landscape of firms in the product market. To address this issue, I adopt a dynamic global text-based industry classification approach based on a unique text dataset from Refinitiv. This innovative methodology allows for the quantification of pairwise product relatedness, surpassing traditional industry classifications and enabling the creation of homogeneous groups, particularly in emerging economies. The chapter concludes by emphasizing the fresh avenues this technique opens for exploring competition in global markets.

Properly grouping firms into industries is a crucial area of research in finance and industrial organization. Traditional industry classifications, such as the SIC, NAICS, and GICS, have limitations that restrict researchers' ability to study product competition and capture changes within industries. These classifications fail to account for cross-country variations in industry homogeneity and lack granularity in identifying emerging industries. To address these limitations, a global network method based on business descriptions can be employed. This method utilizes text data and similarity measures to create a more accurate and granular classification system. It outperforms traditional classifications in explaining firm-level variables, particularly in global and developing markets. The network method also allows for the identification of foreign competitors and provides a comprehensive understanding of competition beyond binary measures like tariffs.

Furthermore, the global network method based on text data has practical applications. For example, it can be used to evaluate whether *Altasia*, a group of economically integrated Asian countries and regions, can replace China in the global supply chain.

By measuring horizontal relatedness of products across countries, the method can identify the best alternative to China for specific industries. Additionally, the network method can effectively measure foreign competition by calculating pairwise similarities. Unlike traditional measures like tariffs, this approach captures the magnitude and extent of competition, enabling a more nuanced analysis of its impact on firms and industries. This method offers a promising avenue for investigating the relationship between competition, including foreign competition, and vertical integration.

In summary, the global network method based on text data presents a more accurate and granular approach to industry classification. It overcomes the limitations of traditional classifications, performs well in global and developing markets, and provides insights into the dynamics of competition and industry changes. This method has practical applications in assessing the potential of alternative regions in the global supply chain and measuring foreign competition beyond binary measures. By leveraging text data and similarity measures, researchers and policymakers can gain a deeper understanding of industries and make more informed decisions.

In Chapter 2, I focus on constructing a measure of the pairwise relatedness of firms' products using textual analysis of their business descriptions. The objective is to investigate whether asset complementarity plays a significant role in mergers and acquisitions (M&A) in the global market. The findings reveal that high similarity between merging firms' products can enhance the profitability of acquiring firms, especially when operating in highly competitive markets. However, the study does not observe improvements in cost reduction, sales growth, or potential product differentiation, indicating that related mergers may not generate asset complementarity.

This chapter explores the impact of asset complementarity on global mergers and acquisitions (M&A) by analyzing a sample of transactions from 2006 to 2018. It utilizes various methodologies, including text-based measures and event studies, to shed light on the potential benefits of asset complementarity, such as introducing new products and enhancing competitive advantage. Findings indicate that, contrary to some previous research, the value creation in mergers may not necessarily come from asset complementarity but from other synergistic benefits, especially when firms operate in highly competitive markets.

Delving further into the implications of asset complementarity in cross-border and domestic transactions, this research suggests that asset complementarity might not be

the primary driver of value creation in these deals. Rather, other types of synergies could potentially be more significant in generating value. Intriguingly, the study notes a negative impact on the long-term performance of business-similar mergers within the U.S., which contradicts the findings of Hoberg and Phillips (2010). This discrepancy emphasizes the importance of considering the context and timeframe of specific M&A activities when analyzing their potential outcomes.

The findings derived from this study present valuable insights that could be instrumental for investors, regulators, and practitioners in the M&A field. Although the results challenge some traditional beliefs regarding asset complementarity in mergers, they simultaneously offer a nuanced perspective on global M&A dynamics. It highlights the complex factors at play in such transactions, including the significance of business similarity, the role of competitive markets, and the distinct effects of public and private targets. Overall, the study underscores the importance of a multi-faceted approach to understanding the mechanisms of value creation in M&A transactions.

Chapter 3 addresses the impact of news on security prices across countries and languages, a topic that has received limited large-scale research. Overcoming the barrier of analyzing news content in different languages, it examines how news in several languages influences returns on stock market indices in various countries. By utilizing a comprehensive dataset of over 270 million online news articles worldwide and employing novel multi-lingual text processing technology, the study assesses the value of news relevant to each country. A "news intensity" variable is developed to evaluate its incremental contribution to explaining stock market index volatility. The results demonstrate that news coverage in different languages has a significant impact on stock markets globally. Greater volumes of relevant news tend to predict a reduction in volatility, suggesting that news coverage serves to resolve uncertainty rather than generate it. The interpretation of results across languages proves more challenging, with languages other than English or the local language frequently exerting a noteworthy influence. Overall, the study highlights the potential and difficulties associated with working with textual big data in an investment context.

To be more detailed, understanding the relationship between news coverage and asset prices is crucial for investors, regulators, and researchers. Previous studies have shown that news coverage influences investor attention, price discovery, and trading volume. However, there has been limited cross-country and cross-language analysis in this area. To address this gap, in this chapter, we utilized a unique dataset of millions

of online news stories in five different languages. By employing a novel textual analysis technology called semantic fingerprinting, we quantified the volume of news relevant to each country and examined its impact on stock market volatility.

The efficient markets hypothesis suggests that asset prices should fully and rapidly reflect public information. Empirical evidence supports the idea that news content affects asset prices, but not always fully and instantly. Moreover, news coverage itself, even without new information, has been found to impact asset prices by directing investor attention. The study of this chapter adds to the existing literature by investigating the influence of news coverage across multiple languages and countries. By analyzing a vast dataset of over 270 million news articles, we were able to predict stock market volatility based on the intensity of news coverage. Interestingly, the findings suggest that greater news intensity tends to have a calming effect on investors and markets, and the impact of news language is not always straightforward.

In conclusion, this chapter contributes to the understanding of the relationship between news coverage and asset prices by expanding the analysis to include indirect, global news coverage across different languages. By employing semantic fingerprinting technology, we were able to measure the volume of news relevant to each country and its impact on stock market volatility. The results demonstrate the significance of news intensity in forecasting stock index volatility and indicate that greater news coverage is generally associated with a calming effect on markets. The study also highlights the importance of considering non-English news sources in the increasingly interconnected global economy. However, the richness of textual big data poses challenges in interpretation, emphasizing the need for further research in this area.

References

[1] Ash, E. and Hansen, S., 2023. Text algorithms in economics. *Centre for Economic Policy Research.*

[2] Bandiera, O., Prat, A., Hansen, S. and Sadun, R., 2020. CEO behavior and firm performance. *Journal of Political Economy*, 128(4), pp.1325-1369.

[3] Bertrand, M., Bombardini, M., Fisman, R., Hackinen, B. and Trebbi, F., 2021. Hall of mirrors: Corporate philanthropy and strategic advocacy. *The Quarterly Journal of Economics*, 136(4), pp.2413-2465.

[4] Gentzkow, M., Kelly, B. and Taddy, M., 2019. Text as data. *Journal of Economic Literature*, 57(3), pp.535-574.

[5] Gentzkow, M., Shapiro, J.M. and Taddy, M., 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4), pp.1307-1340.

[6] Hoberg, G. and Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10), pp.3773-3811.

[7] Hoberg, G. and Phillips, G., 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), pp.1423-1465.

[8] Lang, M. and Stice-Lawrence, L., 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3), pp.110-135.

[9] Shapiro, A.H. and Wilson, D.J., 2022. Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *The Review of Economic Studies*, 89(5), pp.2768-2805.

Chapter 1: Global text-based industry classification*

Xiaojuan LIU†

March 2023

Abstract

Accurately categorizing firms into industries is crucial for conducting research in finance and industrial organization. Standard industry classifications, such as SIC, offer limited information regarding the competitive landscape of firms in the product market. In line with Hoberg and Phillips' (2016) approach, I develop a dynamic global text-based industry classification by leveraging a unique text dataset from Refinitiv. One key benefit of this new methodology is its ability to quantify pairwise product relatedness, which surpasses standard industry classifications in generating homogeneous groups, especially in emerging economies. As a result, this innovative technique offers fresh avenues for exploring competition on global markets.

Keywords: Industry classification, Competition, Textual analysis

*I am very grateful for the guidance, support, and invaluable suggestions from Catherine Casamatta and David Stolin. Special thanks are due to Gerard Hoberg for numerous conversations and suggestions, as well as for hosting me at the University of Southern California as a visiting Ph.D. scholar. I also thank Patrick Coen, Christophe Bisière, Matthieu Bouvard, Sophie Moinas, Dakang Huang, Bao Li, and PhD Workshop in Finance participants at TSM and TSE.

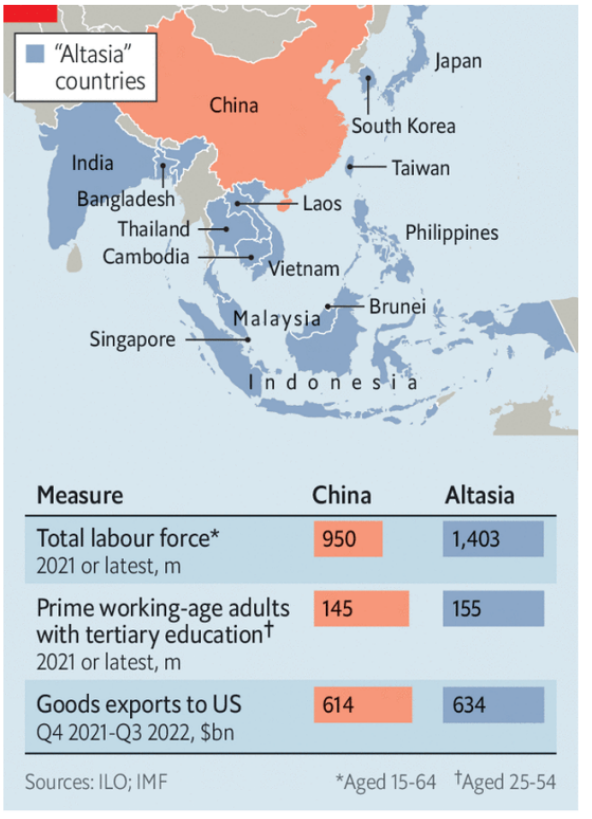
†University Toulouse 1 Capitole, xiaojuan.liu@tsm-education.fr

1 Introduction

Properly grouping firms into industries is a crucial area of research in finance and industrial organization. For instance, researchers must determine if two firms belong to the same industry when studying the effects of horizontal M&A (Hoberg and Phillips, 2010). Industry classifications are also used to inform public policy decisions, with policymakers relying on industry classification data to identify growth and decline in particular regions, and to develop policies to support or restructure industries accordingly. Currently, most studies employ standard industry classifications to meet these objectives. The Standard Industry Classification (SIC), which has been available since 1937, is one commonly used system. The U.S. government developed the North American Industry Classification System (NAICS) in 1997 to replace the SIC system. Additionally, companies offer private industry classifications, such as the Global Industry Classification Standard (GICS), developed by MSCI and S&P Dow Jones Indices. These traditional classifications primarily rely on the production function as the primary criterion to select competitors and assign firms to pre-defined industries.

The standard industry classifications have some limitations which restrict researchers' ability to learn about competition on the product market. For instance, it fails to capture the homogeneity of industries across countries, which has become a critical concern in the current era of globalization and deglobalization. The US-China trade war and the COVID-19 pandemic have forced foreign manufacturers to move out of China, and many Asian countries are emerging as preferred alternatives for relocation. In February 2023, The Economist introduced the concept of *Altasia*, which refers to the 14 economically integrated Asian countries and regions that may replace China's position. *Altasia* stretches from Japan's Hokkaido to Gujarat, northwest India, passing through South Korea, Taiwan, the Philippines, Indonesia, Singapore, Malaysia, Thailand, Vietnam, Cambodia, and Bangladesh. However, the extent to which *Altasia* can substitute China's position in the global supply chain is a matter of dispute. Traditional industry classification systems cannot provide a solution to this problem as they fail to account for cross-country variations in industry homogeneity, which is critical in the current scenario of shifting supply chains.

In addition to the limitations discussed above, traditional industry classifications are unable to provide timely updates on changes within an industry or the emergence of new industries. For instance, the artificial intelligence industry, which is rapidly evolving, is classified under the commercial physical and biological research industry. Similarly, the Fintech sector, which has gained significant importance in recent years, is not included in any specific industry category and is instead classified under either



Source: The Economist

the finance or technology industry. This lack of granularity in traditional industry classifications hampers accurate analysis and policy-making.

To improve the industry classification method in product space, I develop a global network method using a unique text dataset from Thomson Reuters (now Refinitiv) in the spirit of the method of Hoberg and Phillips (2016). The primary objective of this project is to create a text-based industry classification system that can accurately group firms based on their business descriptions. To achieve this, the first step involves vectorizing the text and computing the similarity between any pair of vectors, which ranges from 0 to 1. Firms with high similarity are classified in the same sector, allowing for a more granular classification system than traditional industry classifications. Following Hoberg and Phillips (2016), I apply two algorithms in this project. The first algorithm generates a fixed number of 300 industries based on pairwise similarities, mimicking the number of industries in the SIC-3 code.[1] The second algorithm is a network approach that identifies 2% of competitorship based on a business similarity threshold above which firms will be regarded as competitors. This approach is inspired by the 2% competitorship rule employed in the SIC-3 classification system. The

[1]Hoberg and Phillips (2016) use the same approach to compare their 10-K text-based classification to the SIC system.

threshold identified in this project is 21.52%, which is comparable to the value found in Hoberg and Phillips (2016), which is 21.32%.

In this chapter, I compare the effectiveness of the global text-based industry classification method with the standard industry classifications (SIC, NAICS, and GICS) and the 10-K-based text network method (TNIC) established by Hoberg and Phillips (2016). To validate the performance of these methods, I employ the approach developed by Bhojraj et al (2003), which is widely used in this field (see Hrazdil, Trottier, and Zhang, 2013; Kaustia and Rantala, 2021). The primary objective of this approach is to evaluate the ability of industry classifications to group homogeneous firms. One way to measure homogeneity is to examine how similarly the market values the crucial accounting and financial metrics of firms in the same industry. The test variables I use are the return on assets (ROA), the return on equity (ROE), the Market-to-Book ratio, the Price-to-Book ratio, the Return on Net Operating Assets, the Assets Turnover, and the Leverage. For each industry classification scheme, I form industry portfolios and examine to what extent the mean industry variables can explain firm-level variables. In the regression analysis, I consider only the adjusted R^2, as a high adjusted R^2 indicates a small distance between a firm's financial metrics relative to that of industry averages.

The existing literature primarily focuses on industry classification comparison in the US market. Bhojraj et al (2003) and Hrazdil, Trottier, and Zhang (2013) compare the SIC, NAICS, GICS, and Fama-French industry classifications in the U.S. market. Similarly, Hoberg and Phillips (2016) compare TNIC with SIC and NAICS in the U.S. market, while Kaustia and Rantala (2021) compare analysts-based methods with SIC, NAICS, GICS, Fama-French, and TNIC in the U.S. market. However, to the best of my knowledge, no study has yet verified the effectiveness of industry classification schemes in global markets. In my research, I have developed a text-based method and analyzed its performance on global markets, developed markets (excluding the U.S.), developing markets, and the U.S. market.

This analysis reveals that the network method is highly effective in explaining five variables on global markets, except for Assets Turnover and Leverage. The performance of the network method on the developed markets is quite similar to that on global markets. Moreover, the network method demonstrates remarkable performance in developing markets, where it outperforms almost all other classification schemes. In contrast, the network method underperforms in the U.S. market compared to GICS and NAICS codes. This finding is consistent with the results reported by Kaustia and Rantala (2021) which show the GICS outperforms the TNIC. Besides, Bhojraj et al (2003), and Hrazdil, Trottier, and Zhang (2013) also suggest the GICS is the best-

performing classification scheme among all traditional systems in the U.S. market. Additionally, when comparing the network method based on short business descriptions from Refinitiv (henceforth"TRBD") with the TNIC scheme for U.S. firms, no significant difference between the two methods was detected.

Furthermore, Bhojraj et al (2003) have suggested that firm size can impact the effectiveness of industry classifications, with larger firms performing better than smaller firms. To determine whether this holds true on global markets, I divided my dataset into three groups based on their market value. Interestingly, I find that the network method based on TRBD is most effective for small firms, followed by large firms and then mid-cap firms. This finding is novel and adds to the existing literature on industry classifications.

In addition, the network method based on TRBD allows me to easily distinguish between domestic peers and foreign competitors. To investigate the impact of foreign competitors on the analysis, I ran regressions comparing the adjusted R^2 of models with and without foreign competitors for the ROA variable. However, I did not observe a significant improvement in adjusted R^2. Nevertheless, the coefficient of the average foreign peers' ROA remained positively significant at a 99% confidence level, even when I further divide foreign competitors into those from developed and developing markets. This result highlights the importance of considering foreign competitors in such analyses, and supports existing literature on the topic.

My method has numerous applications, one of which is answering the question of whether *Altasia* can replace China by measuring the horizontal relatedness of products from China and *Altasia* countries. For example, countries like South Korea, Japan, Taiwan, and China are all major semiconductor makers on the global market. With the current deterioration of the relationship between the U.S. and China, it is possible that these other countries and areas could replace China in exporting semiconductors to other countries. By measuring the degree of product homogeneity across countries, we can determine which country is the best alternative to China for semiconductor production.

The network method based on TRBD also has another important application - measuring foreign competition. Typically, changes in tariffs are used as a measure of foreign competition in empirical research (See Srinivasan, 2020). However, this approach has two major limitations. Firstly, it only measures the industries affected by the tariff changes and provides little information on non-vertically related industries. Secondly, the binary nature of the tariff measure limits its ability to capture the magnitude of the change. For instance, this approach cannot verify a U-shaped relationship between competition and vertical integration, as suggested by Aghion et al. (2006). The global

network method overcomes these limitations by calculating the foreign competition for any given firm by summing up all pairwise similarities. This provides a more comprehensive and nuanced understanding of the relationship between competition (including foreign competition) and vertical integration. Therefore, the global network method offers a promising approach to investigating the impact of foreign competition on firms and industries beyond the binary nature of tariff measures.

The rest of the chapter is organized into 4 parts, Section 2 presents the extant industry classifications as well as their limitations. Section 3 introduces the data and the machine learning algorithm applied to form the global network industry classification. Section 4 shows the econometric method and empirical results. Section 5 concludes.

2 Background information

In this section, I introduce the existing industry classification systems including their classification criteria and development. I point out the limitations they bear and how I improve these drawbacks by developing a global text-based industry classification scheme.

2.1 Introduction of the classical industry classifications

2.1.1 Standard industry classification

Standard industry classification (SIC) is the oldest one among all the industry classification systems and it was constructed by Interdepartmental Conference on Industrial Classification in 1937. The goal was "to develop a plan of classification of various types of statistical data by industries and to promote the general adoption of such classification as the standard classification of the Federal Government." (Pearce, 1957). At that time, there was no commonly used industry classification system, and different agencies applied different rules and made the comparisons of industries difficult. The system first focused on the manufacturing industries and later extended to the nonmanufacturing ones. Four subcategories represented by digits from 1 to 4 are under the SIC system. Given that the SIC scheme has such a long system, it has been used in numerous research studies. Data providers such as Compustat and Datastream provide the SIC codes in their databases.

2.1.2 North American Industry Classification System

North American industry classification System (NAICS) was established by the U.S. government in 1997 to replace the SIC codes. It was a joint work of the United States, Canada, and Mexico under the historical background of a free trade agreement for North America. This system is revised every five years to ensure its relatedness to economic activities. Similar to the SIC system, the NAICS also uses different digits to represent industry groups from digits 2 to 6. There are 5 subcategories under this structure. The NAICS is a production-oriented framework that emphasizes firms' similarity in inputs used to produce goods or services rather than outputs when forming industries.[2] For the years before 1997, the NAICS was backfilled in Compustat. One innovation of this scheme is that it makes the industry classifications in three countries comparable and make comparative research easier to conduct. Empirical work such as Krishnan and Press (2003) suggests the NAICS scheme generates more homogeneous groups compared to the SIC system.

2.1.3 Global Industry Classification Standard

Global industry classification Standard (GICS) is a private system built by MSCI and S&P Dow Jones Indices and was designed to fit the needs of financial professionals in 1999. The GICS system has 4 subclassifications and it uses digits ranging from 2 to 8 to designate the industries, which are GICS-2, GICS-4, GICS-6, and GICS-8 respectively. Figure 1 shows the hierarchical structure of the GICS system.[3] The revenues of firms are applied by the GICS system to define firms' principal business activity. Besides, the GICS also takes earnings and market perception into consideration to determine firms' industries. Practitioners such as asset managers have widely adopted the GICS for financial research and the GICS can also apply to global companies. The GICS codes in Compustat before 1999 are backfilled and Datastream only provides the latest information on firms' GICS industries. In research that compares different industry classifications (Bhojraj et al, 2003; Hrazdil, Trottier and Zhang, 2013), academics find the GICS performs well relative to the SIC and the NAICS.

[2]See 2022 NAICS Manual: "NAICS is an industry classification system that groups establishments into industries based on the similarity of their production processes."

[3]Resource: https://www.msci.com/our-solutions/indexes/gics.

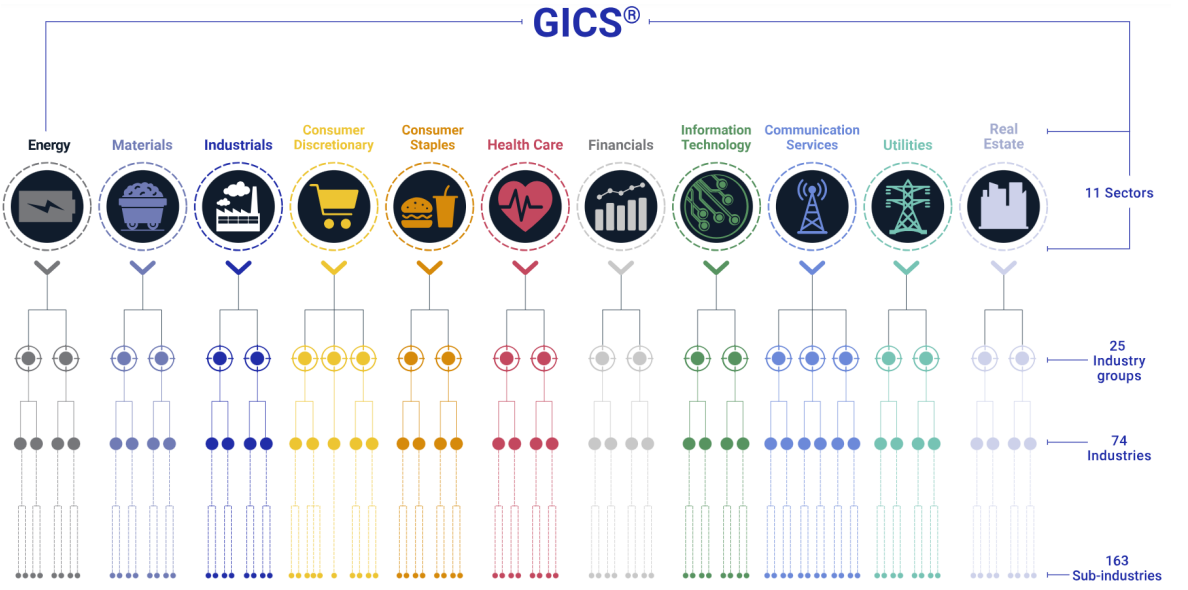


Figure 1: The four-tiered GICS structure

Source: https://www.msci.com/our-solutions/indexes/gics

2.2 Text-based industry classification

2.2.1 Limitations of the standard industry classifications

The existing industry classifications have several apparent disadvantages. One example is the inconsistency in the SIC codes recorded by Compustat and CRSP databases. The work conducted by Guenther and Rosman (1994) suggests 38% of the firms have different SIC-2 codes when they compare Compustat and CRSP. This problem can bias empirical results when researchers use different databases. Besides, there are other disputes over the traditional industry classifications documented in the literature.

First, Hoberg and Phillips (2016) show firms can be more similar in the product space than the SIC codes suggest. One example can be the newspaper industry and the radio industry. Both industries produce and collect information and transfer the daily news to the audience. However, these two industries have very different SIC-3 codes. The newspaper publishing and printing industry yet has a SIC-3 of 271. The radio and broadcasting stations industry has a SIC-3 of 483. When one considers competition in the media industry, newspapers and radio should be taken as rivals. Newspapers and radio both aim to capture and retain the audience's attention. They provide news, entertainment, and other content to engage and inform their audiences. As a result, they are in direct competition for consumers' limited attention spans. If a person spends more time listening to the radio, they may spend less time reading

newspapers, and vice versa.

Second, the standard industry classifications cannot capture the distance between firms in the product space, neither within one industry nor across industries. A standard industry classification is a binary system that informs the user about whether a given firm is in one industry or not. The appearance of two firms in the same industry defines competitors. This criterion causes several inconveniences. i) It fails to capture the business changes of industrial evolution. In the past, petroleum products were the primary energy sources used by cars. Nowadays, other energy sources such as electricity are also consumed by cars. This innovation is not captured by the standard industry classification but is important, for example, from the perspective of climate change. ii) It fails to measure business relatedness among industries. The market observes a negative correlation between the sales of smartphones and cameras. If the researcher wants to investigate the causality behind this phenomenon, one important step is to measure the business relatedness between these two industries. Both smartphones and cameras can take photos when these two industries specialize in different fields. Does this business relatedness cause the fall in sales of cameras? The binary system of standard industry classification cannot answer this question. iii) It fails to predict firms' entries into and exits from industries. The business of a given firm can be dynamic and one firm may move from one industry to another industry. The standard industry classification only sees the result when a firm finally changes the industry. A more dynamic measure is necessary if we want to predict which industry a given firm may enter or exit.

Third, standard industry classifications cannot respond to new industries in a timely manner. A famous example can be technology and web-based industry. As Hoberg and Phillips (2016) document, this industry was classified into the business service industry when it emerged in the early 2000s. A recent example is Fintech firms. There is no industry includes all Fintech firms and these firms are classified into either the finance industry or the technology industry.

Fourth, transitivity is imposed by the standard industry classifications. Assume there are three firms on the market, firm A, firm B, and firm C. When A and B are classified as competitors, as are A and C, A and B must be rivals according to a binary system.

2.2.2 Text-based Network Industry Classifications

To improve the standard industry classifications, Hoberg and Phillips (2016) use Item 1 of 10-K documents to build a dynamic textual industry classification. One important difference between standard industry classifications and textual industry

classification is the criterion applied to classify the firms. The text-based method mainly focuses on the final products or services sold on the market. This method provides researchers with a new dimension to know how firms compete in the product space.[4] Another characteristic of this method is that it can measure the similarity of business for any given pair of firms. This advantage allows researchers to capture the relatedness of peers among and across industries. It is important because products can be very similar in one industry such as the consumer electronics industry but quite different in another industry, say the automotive industry. In the consumer electronics industry, products are often very similar in terms of their basic features and functions. For example, smartphones from different manufacturers may have slightly different designs or user interfaces, but they all perform the same basic functions. Whereas in the automotive industry, products can vary significantly in terms of features, performance, and design. In addition, Hoberg and Phillips (2016) show that the text-based method can react to the evolution of new industries in a timely manner and better capture the impact of a given shock compared to standard industry classifications. Furthermore, the text-based method is updated annually and relaxes the transitivity of competitorship.

Their work has been applied widely in finance, economics and industrial organization domains such as: How does policy uncertainty affect mergers and acquisitions (Nonaime, Gulen, and Ion, 2018); Measure the valuable of the FinTech innovation (Chen, Wu, and Yang, 2019); Detect the relation between intangible capital and the investment-q (Peter and Taylor, 2017); Model CSR as an investment to increase product differentiation that allows firms to benefit from higher profit margins (Albuquerque, Koskinen, and Zhang, 2019); Show the evidence of institutional investors value and demand climate risk disclosures (Ilhan et al., 2023); The relationship between firm investment and peers' investment (Bustamante and Fresard, 2021); Test whether a private firm's decision to go public affects the IPO decisions of its competitors(Aghamolla and Thakor, 2022); Propose a novel firm-level measure of cybersecurity risk for all U.S.-listed firms based on textual analysis(Florackis et al., 2023); Using machine learning score the five corporate cultural values of innovation, integrity, quality, respect, and teamwork (Li et al., 2021); Examining investment banks' choice of peers in comparable companies analysis in M&As (Eaton et al., 2022); Analysis the information flow driven by investment opportunity based on a pairwise measure (Bernard, Blackburne and Thornock, 2020).

However, Hoberg and Phillips (2016) concentrate on the U.S. market. As global-

[4]Competition also happens in other dimensions, for example in the technology space, see Bloom, Schankerman, and Reenen (2013).

ization matters more and more in recent years, how firms compete on international markets is a crucial subject for both academic research and industry. Knowing the product and industry evolution on global markets will give firms advantages to create their own market or innovation strategies. Competition in the domestic and foreign markets also attracts the attention of academic research. If we want to learn more about competition in Artificial intelligence, smartphone, and the 5G, we cannot ignore Asian and European markets.

In this chapter, I use a unique text dataset from Thomson Reuters (Refinitiv) which covers 96% of public firms based on the market value on the global market to build a dynamic industry classification. With this methodology, I can measure the homogeneity of products and industries across countries. Separating foreign competition from domestic competition is facilitated by the network method. Traditional empirical research mainly uses tariff changes to gauge the modification of foreign competition which loses information on the magnitude of the change. My method can fix this problem easily.

2.3 Other industry classifications

There are other available industry classification systems that I don't test in the chapter.

2.3.1 Fama-French industry classification

The Fama-French industry classification is a widely used method for categorizing companies into industry groups based on their business activities. It was developed by Eugene Fama and Kenneth French and is commonly employed in financial and academic research. One application of the Fama-French industry classification is in estimating the cost of equity, which is an important metric used in financial analysis.

Fama and French (1997) assign firms to 48 industries using SIC-4 codes. They don't aim to build a new industry classification structure but to better solve the cost of equity problem. In several industry classifications comparison research, the Fama-French industry classification always underperforms compared to other systems such as GICS.[5] The Fama-French industries data can be obtained through the internet and the dataset starts from 1926.[6] This industry classification is too broad for the current study.

[5]Ssee Bhojraj et al. (2003), Hrazdil, Trottier and Zhang (2013), and Kaustia and Rantala (2021).

[6]See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_48_ind_porthtml

2.3.2 Industry Classification Benchmark

Industry Classification Benchmark (ICB) is a private industry classification structure constructed by FTSE Russell. The ICB system uses the source of a firm's revenue or the majority revenue as the benchmark to classify firms into groups. The documents employed to detect the revenue are the audited accounts and directors' reports. When the revenue information is not available, the ICB scheme will use the business description information from the annual report to define firms' groups or the business summary from the company's website. Academics rarely apply the ICB for the comparison of the industry classification efficacy. The reason may be limited data availability. Datastream provides information on firms' ICB industries but the data is the latest value.

The method employed in this approach uses similar techniques to the text-based method in grouping firms. However, there are notable differences between the ICB and the text-based method based on TRBD. These differences can be summarized as follows: i) Business description: In the ICB industry classification, business description serves as the final criterion for forming industries. On the other hand, the text-based method relies solely on business description to determine industry classification. ii) Revenue-based classification: ICB primarily utilizes revenue, or the majority revenue, to establish industry categories. To illustrate, let's consider Apple Inc. as an example. Apple Inc.'s annual report for the year 2022. The report reveals that the majority of Apple's revenue is derived from iPhone, Mac, and iPad, while the second-largest revenue source is iCloud and other products. According to ICB rules, only the first category source of revenue would be considered when forming an industry. In contrast, the text-based method takes into account all the products available in the market when determining industry classification. By highlighting these distinctions, it becomes evident that the ICB and the text-based method differ in their treatment of business descriptions and revenue sources. While ICB places greater emphasis on revenue and limits industry formation to specific revenue sources, the text-based method considers a broader range of products to establish industries.

2.3.3 The Refinitiv Business Classification

The Refinitiv Business Classification (TRBC) is an industry classification system developed by Refinitiv. It is a market-based classification scheme in which firms will be grouped into industries based on the market they operate in rather than the products or services they provide. The TRBC scheme emphasizes the similar market characteristics shared by enterprises when forming groups. Under the TRBC, airlines' catering services

are classified as airport services rather than restaurants. The analysts of Refinitiv collect financial reports as the main resources for the classification process. The TRBC industries are available on Eikon which is a product of Refinitiv. Similar to Datastream, Refinitiv only provides the latest value of the TRBC data.

2.3.4 Analyst-based method

Researchers believe the coverage of analysts of firms echoes the other similarities among firms such as customer segment and business model. Earlier, Ramnath (2002) defines analyst-based industry groups in which firms share at least 5 common analysts with peers. This new fashion is attractive but leaves some firms to remain unclassified. Later, Kaustia and Rantala (2021) compare analysts' actual coverage choices to simulated random choices to identify peer groups of firms based on common analyst coverage. This method is self-organizing and does not specify the size or granularity of the peer groups ex-ante. Similar to the TNIC, the analyst-based method is dynamic and can respond rapidly to changes in firms and industries, and this method also relaxes the transitivity of competitor relationships. The advantage of the analyst-based method is apparent. However, this method fails to provide insight into the criteria and process used by analysts to select groups of firms considered as competitors. Additionally, it lacks an objective measure of competition, leaving unanswered questions about how competitive relationships between firms are determined. Besides, all related analysis is restricted to U.S. firms.

3 Data and algorithm

In this section, I will introduce the data I apply in this chapter and the machine learning algorithm I use to form the industries.

3.1 Data description

3.1.1 Text data description

The data employed in this chapter is the short business descriptions from Refinitiv ("TRBD") from the year 2006 to 2021.[7] These descriptions are available globally. An extensively used data source in textual finance is 10-K filings in the U.S.[8] Item 1 of 10-K which mainly describes the business of a firm has been used for the industry

[7]The reason why the data starts from the year 2006 is that the coverage of data becomes globally comprehensive from this year.

[8]See Hoberg and Philips, 2010 and Hoberg and Philips,2016 for more information.

classification work (See Hoberg and Philips, 2016). Given 10-K filings are restricted to U.S. firms, the industry classification relying on 10-K cannot be extended to other countries. To address this concern, I use a new database that provides the business description of global public firms to build a global text-based industry classification scheme. The primary source of this business description is the financial reports that firms filed, which may be annual, semi-annual, or quarterly. Refinitiv cooperates with over 130 disclosure authorities worldwide to collect these documents. The institution has operations working in different countries such as the U.S. and China to ensure the accurate translation of business information provided by firms. To timely monitor the business changes of firms, Refinitiv also focuses on the external M&As taken by firms recorded by press releases and other media reports.[9] Here is an example of TRBD of Microsoft company. For firms that operate in different countries, TRBD also provides information on these countries.

"*Microsoft Corporation is a technology company. The Company develops and supports software, services, devices, and solutions. Its segments include Productivity and Business Processes, Intelligent Cloud, and More Personal Computing. The Productivity and Business Processes segment consists of products and services in its portfolio of productivity, communication, and information services, spanning a variety of devices and platforms. This segment includes Office Consumer, LinkedIn, dynamics business solutions, and Office Commercial. The Intelligent Cloud segment consists of public, private, and hybrid server products and cloud services that can power modern businesses and developers. This segment includes server products and cloud services, and enterprise services. The More Personal Computing segment consists of products and services that put customers at the center of the experience with its technology. This segment includes Windows, devices, gaming, and search and news advertising.*"

We can see from this TRBD that it covers mainly the information on the products sold on the market including the products' names. After the comparison of TRBD and Item 1 of 10-K, I find there are three main differences between the two datasets: i) The TRBD is much shorter than Item 1 in 10-K filings on average.[10] For example, Hoberg and Philips (2010) remove the firms whose product description contains fewer than 1000 characters. Regarding TRBD, the average number of characters in each document is 122. As a result, TRBD is quite brief and contains fewer uninformative words. ii) 10-K filings are updated annually. Given the fact that TRBD comes from the financial reports, TRBD updates at least once a year. Besides, Refinitiv also brings up-to-date firms' business information when there are transactions in which firms get

[9]The information is from the internal document of Refinitiv.

[10]The length of Item 1 of 10-K varies from less than 1 page to more than 50 pages.

involved. Thus, TRBD updates more frequently than 10-K filings. To verify, I select the Dow Jones firms and check the frequency at which TRBD changes. The result shows that these firms' TRBD is updated more than once a year for several firms. To ensure TRBD's changes are not merely cosmetic ones, I randomly select 100 firms whose TRBD has been updated. I read the ex-ante and ex-post TRBD and I find the updates are related to the firms' business changes. iii) TRBD is prepared by the same third party rather than by the firms themselves, which makes TRBD much more homogenous.

On average, 48,280 firms have available TRBD each year and there is an increasing trend of the number of firms from the year 2006 to 2021. The coverage is 95% based on the market value of firms across the years. The coverage based on the number of firms each year is 85% on average. Table 1 shows the details of the coverage of TRBD.[11]

Year	BD	Total firms	Coverage based on number of firms	Coverage based on MV
2006	37457	51355	73%	91%
2007	42414	53627	79%	95%
2008	44754	54299	82%	94%
2009	46241	56694	82%	95%
2010	46515	55219	84%	95%
2011	47111	55721	85%	95%
2012	47531	55380	86%	96%
2013	47198	54624	86%	96%
2014	47220	54899	86%	97%
2015	47685	55121	87%	97%
2016	49665	54787	91%	97%
2017	50234	54728	92%	98%
2018	52489	60990	86%	94%
2019	55451	64496	86%	95%
2020	53851	62940	86%	94%
2021	56664	59569	95%	100%
	48280	56528	85%	96%

Table1: Number of global public firms for each year

TRBD can be downloaded directly from Eikon which is a terminal of Refinitiv. However, the business description is the latest information and the historical data is

[11]From the year 2018, there is a drop in the estimated coverage. This is from the switch of the database where I collect the number of total global public firms. From the year 2006 to 2018, I collect the firms from Datastream and later on, I download the firms from Eikon.

unobtainable from Eikon. Starting in 2018, I regularly downloaded all available business descriptions from Refinitiv. Additionally, I have been able to purchase historical TRBD from Refinitiv company for the years 2006 to 2017. One important concern with historical data is the potential survivorship bias if delisted firms can be missing from the dataset. If the dataset does not cover the dead firms well, it will lose much potential in the industry classification work as survivorship bias can invalidate a study's conclusions (See Brown et al.,1992). To investigate the coverage of the delisted firms of this dataset, I download the global universe of equities from both Datastream and Factset from the year 2006 to 2017 and compare them with the historical TRBD. Datastream provides information including market value, public status, ISIN, etc. Based on the empirical analysis, the coverage of currently listed public firms is around 96% on the basis of the market value and the coverage of delisted firms is around 92%. Table 2 shows the details of the coverage for both delisted and listed firms from the year 2006 to 2017.

Year	TRBD of delisted	TOTAL delisted	Coverage of delisted firms based on MV	TRBD of listed	TOTAL listed	Coverage of listed firms based on MV	TOTAL FIRMS
2006	14145	23266	85.07%	23312	28089	92.73%	51355
2007	15246	23051	90.41%	27168	30576	95.81%	53627
2008	15316	22179	94.17%	29438	32120	95.68%	54299
2009	14739	22426	89.44%	31502	34268	95.66%	56694
2010	13299	18811	91.57%	33216	36408	95.67%	55219
2011	12070	17176	92.27%	35041	38545	95.95%	55721
2012	10792	15146	92.78%	36739	40234	96.15%	55380
2013	9292	12966	94.03%	37906	41658	95.93%	54624
2014	7820	11199	95.97%	39400	43700	96.63%	54899
2015	6359	9221	96.44%	41326	45900	97.51%	55121
2016	5001	6940	84.98%	44664	47847	98.08%	54787
2017	3095	4237	99.43%	47139	50491	98.14%	54728
	10598	15552	92.21%	35571	39153	96.16%	54705

Table 2: The coverage of TRBD on delisted and listed firms

From the year 2018 until now, I download the TRBD of global public firms from Eikon each quarter. For example, in 2019, I downloaded the TRBD of all public firms in March (51,309 firms in total), June (51,309 firms in total), September (49,690 firms in total), and December (49,985 firms in total). I combine all the datasets and I drop the duplication when keeping the last observation. As a result, there are 55,451 firms for the year 2019. From these numbers, we can observe that there are 'exits' and 'new entries' within one year. Therefore, the survivorship bias problem is effectively eliminated by updating TRBD quarterly (as I have been doing since 2018) rather than annually.

Another crucial characteristic of this database is the country coverage. For the historical dataset, I use the identifier of firms to get their domestic country information from Eikon.[12] From the year 2018 until now, I obtain the country of headquarters for firms when I download their TRBD. After the investigation, I find the dataset covers 112 countries. Figure 2 shows the top 20 countries with the highest number of firms covered by the database. To the best of my knowledge, TRBD is the only database that provides continuous comprehensive coverage of public firms' business descriptions across countries.[13]

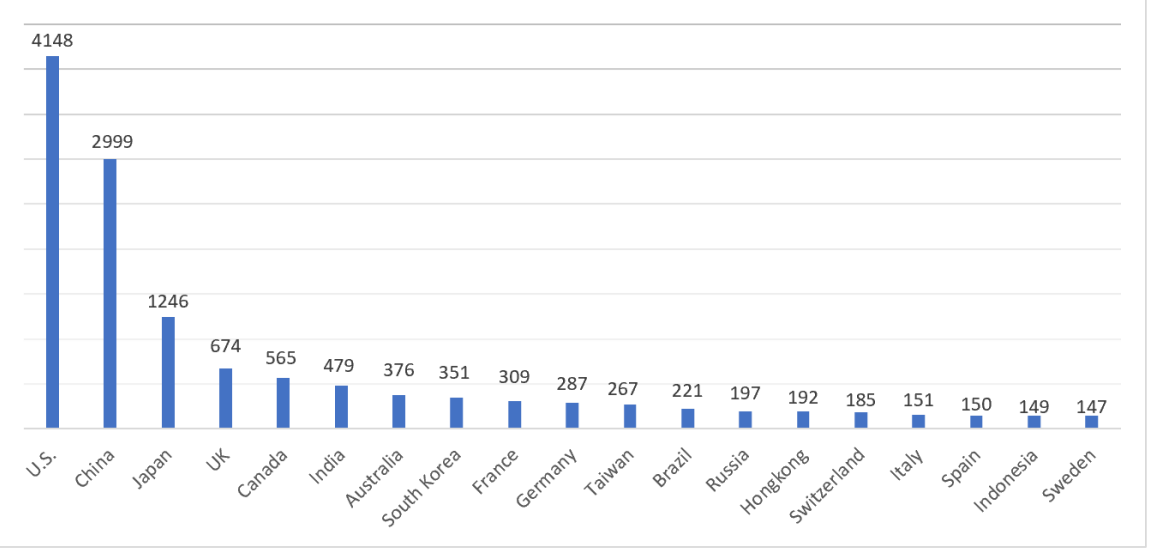


Figure 2: The top 20 countries of TRBD in the year 2006

3.2 Measuring similarities between company descriptions

To form the industries, I first describe how I measure the similarity between texts. The method used in this project is the bag-of-words method which is commonly used by other literature (Hoberg and Phillips, 2016). The first step of the computation is the vectorize the document. To realize it, I need a dictionary for each year as a filter. Here is an example to illustrate. There are 37,457 firms in total in the year 2006. I start the data processing by deleting the sentences of each TRBD when it has words 'acquired', 'merged', or 'sold'. These sentences describe the transactions a given firm had in the past but typically do not have useful keywords for textual analysis of their

[12]This function does not work well after the update of the Eikon in 2021.

[13]Databases such as Capital IQ also provide similar short business descriptions but they do not provide historical business descriptions.

business. I convert all words into lowercase and remove the stop words[14] as well as the string with less than 3 letters[15] in each TRBD. Then I use the tag method and only keep nouns, and proper nouns in all TRBD.[16] I transform all the words of the plural forms to singular forms which reduce the length of my standard vector and makes computations more accurate. Later, I pool all the words that remained in all the 37,457 TRBDs. Hereafter, I discard the words that appear in more than 25% of TRBD which are defined as common words as they are unlikely to be informative. I also remove the words appearing in fewer than 3 different TRBDs as they are too rare to be useful. As a result, there are 10,694 keywords left for the year 2006 and they are used as a dictionary to vectorize the TRBDs in the corresponding year. Figure 3 shows the number of words remaining in the dictionary each year.

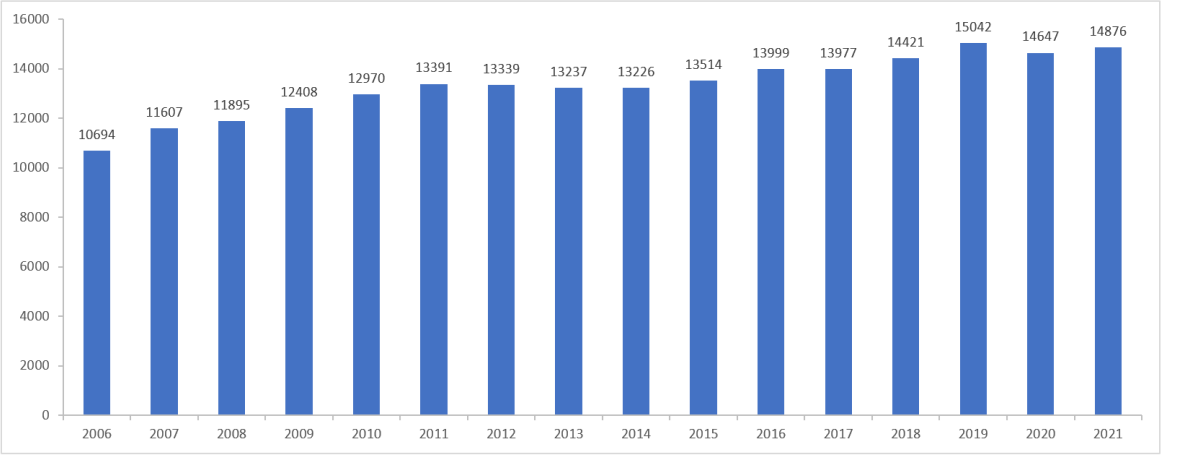


Figure 3: Number of words in the dictionary in each year

With the dictionary, I can vectorize a TRBD. I keep the words of TRBD that are in the corresponding year's dictionary. After the data processing, a small number of TRBDs become empty, and I remove all these empty documents.

Now, I start to vectorize the TRBD and binarize the vector by assigning 1 to the positions where the word exists in both TRBD and the dictionary and 0 for the rest. The binary vector is expected to be sparse. As often discussed in textual analysis literature, cosine similarity helps to gauge the distance of firms in the product space. The formula of cosine similarity for two vectors **A** and **B** is as follows:

[14] I apply the 'StopWords_Generic.txt' developed by Loughran-McDonald, See the website for more details: https://sraf.nd.edu/textual-analysis/stopwords/.

[15] The words with only 1 or 2 letters are not informative of their business.

[16] I also tried to keep adjectives, but the results show most of the adjectives are not very informative about the business. For example, after data processing, some TRBDs only contain adjectives but one is not informed about their business based on these words.

$$\text{similarity } = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are components of vector $\mathbf{A}$ and $\mathbf{B}$ respectively.

3.3 Cluster algorithm

3.3.1 Fixed method

The first method is to form a fixed number of industries. The algorithm is to group firms based on their pairwise similarity score in the business description. Following Hoberg and Phillips (2016), I use single-segment firms to form the industry clusters in the first method. The database which I apply to distinguish single-segment firms from multi-segment firms is Datastream.[17] Datastream provides SIC-4 codes for global firms, and it assigns the SIC-4 codes based on the sales of the segments for a given firm. [18] A given firm can have a maximum of eight SIC-4 codes. The first SIC-4 will be the segment that has the highest sales while the eighth SIC-4 will be the segment that has the lowest sales. I download all eight SIC-4 codes for all firms in my database across the years.[19] Then, I require the firms' second SIC-4 codes to be 'NULL'. This method identifies all the single-segment firms in my sample. Table 3 describes the number of single-segment and multi-segment firms across the years.

From Table 3, we can observe that the percentage of single-segment firms is roughly half that of multi-segments firms. I compare this ratio with the value found in Hoberg and Phillips (2022). I find that there are more single-segment firms in the sample of Hoberg and Phillips (2022). There are three potential reasons: i) The SIC codes in my sample which are used to distinguish the single-segment firms are the latest values. Firms evolve from time to time, a single-segment firm in the year 2006 may become a multi-segments firm in the year 2023. This fact will bias my result. ii) There are more multi-segment firms outside of the U.S. market. iii)Database. I use Datastream to distinguish single-segment firms from multi-segment firms when Hoberg and Phillips (2016) use Compustat.

The next step is to group single-segment firms into a target number of industries. To describe the idea, assume there are 10,000 single-segment firms in the year 2006 and they are labeled from 1 to 10,000. In the beginning, each firm represents an industry.

[17] Compustat is used in Hoberg and Phillips (2016) to find out the single-segment firms.

[18] Datastream documents that if they have no information on the sales of the company, they will use the Worldscope database to assign the SIC codes to firms.

[19] One weakness of the data is that Datastream only provides the latest value of SIC-4 codes. For example, when I download the SIC codes for firms in the year 2006, the data I get is not historical data but the current SIC codes for these firms.

Year	Number of TRBDs	TRBDs after processing	Single-segment firms	% of single-segment firms	Multi-segments firms	% of multi-segments firms
2006	37457	37421	10507	28.08%	26914	71.92%
2007	42414	42379	12377	29.21%	30002	70.79%
2008	44754	44703	13352	29.87%	31351	70.13%
2009	46241	46169	13824	29.94%	32345	70.06%
2010	46515	46463	13417	28.88%	33046	71.12%
2011	47111	47068	15506	32.94%	31562	67.06%
2012	47531	47488	13844	29.15%	33644	70.85%
2013	47198	47160	13774	29.21%	33386	70.79%
2014	47220	47177	13916	29.50%	33261	70.50%
2015	47685	47642	16337	34.29%	31305	65.71%
2016	49665	49574	16188	32.65%	33386	67.35%
2017	50234	50140	16680	33.27%	33460	66.73%
2018	52489	52419	20313	38.75%	32106	61.25%
2019	55451	55378	21209	38.30%	34169	61.70%
2020	53851	53783	20966	38.98%	32817	61.02%
2021	56664	56576	18929	33.46%	37647	66.54%

Table 3: Number of single-segment and multi-segment firms across years.

Then, I compute a matrix of 10,000*10,000 to save the pairwise cosine similarities of these firms. Next, I group the firms with the highest similarity score in the matrix into a new industry. Assuming the highest similarity score is between firm 1 and firm 2, I put the two firms into one industry and label it as 10,001 when deleting firms 1 and 2 from the matrix. At this stage, there are 9,999 industries remaining, and I recompute the cosine similarities among these 9,999 industries to form a new matrix. For the industry 10,001, I will compute the similarities between firm 1 and the remaining of the 9,998 firms as well as that between firm 2 and the 9,998 firms and take the average as the new cosine similarity score for the industry 10,001. The calculation is the same for industries that have more than 2 firms. The steps will repeat until I get the ideal number of groups. The target number of industries in this chapter is set to be 300 as Hoberg and Phillips (2016) find SIC-3 outperforms other versions of the SIC code and it has around 300 industries. To ensure my research is comparable to theirs, I choose the same number of industries for the fixed method.

When 300 industries are ready, I find 5 central firms for each group. The average value of the pairwise similarities between a given company and the other companies within the group is defined as 'centrality'. The central firms are those with the highest centrality for a given industry. Starting from the following year, I group firms into industries based on the average pairwise similarities between a given firm and the central firms in each industry.

In this way, it is guaranteed that the 2006 and 2007 groups will be centered around

Year	TRBDs after processing	Number of firms in the largest industry	% of largest industry	Number of firms in the smallest industry
2006	37421	2270	6.07%	1
2007	42379	2035	4.80%	1
2008	44703	2770	6.20%	5
2009	46169	2479	5.37%	11
2010	46463	1747	3.76%	7
2011	47068	2043	4.34%	10
2012	47488	1514	3.19%	12
2013	47160	1320	2.80%	8
2014	47177	2060	4.37%	11
2015	47642	1488	3.12%	10
2016	49574	1632	3.29%	13
2017	50140	2744	5.47%	8
2018	52419	2969	5.66%	6
2019	55378	3535	6.38%	7
2020	53783	3239	6.02%	6
2021	56576	3990	7.05%	11

Table 4: Number of firms in the largest industries and the smallest industries across years

similar 5 firms, likewise 2007 and 2008 clusters, etc. This ensures continuity. It is also possible these central firms gradually become less central as the industry evolves. Eventually, they can even move to another industry. The industry groups evolve but slowly enough so that they will still be recognizable year to year.

The last step is to assign the multi-segment firms to the groups formed each year. The idea is to find the industry which has the highest similarity score with a given multi-segment firm. I first compute the pairwise similarities between a given multi-segment firm and all the firms in each sector for a given year. Then, I calculate the average pairwise similarities between the given firm and each sector. The sector with the highest average value includes the firm in their group.

Table 4 shows the details of the number of firms in the largest industries and smallest industries. Hoberg and Phillips (2016) document that there are numerous single-firm industries and industries with a huge number of firms for the fixed method. My result presents a different trend that there are fewer single-firm industries and no huge industry. The largest industry of the fixed method in Hoberg and Phillips (2016) contains 20% of firms and that ratio in my sample is around 5% on average, this number is similar to SIC-3 which is around 6%. Figure 4 shows the distribution of the number of firms in each industry of the fixed method and SIC-3 on the global

market in the year 2020. One can observe the distributions of two systems are quite similar.

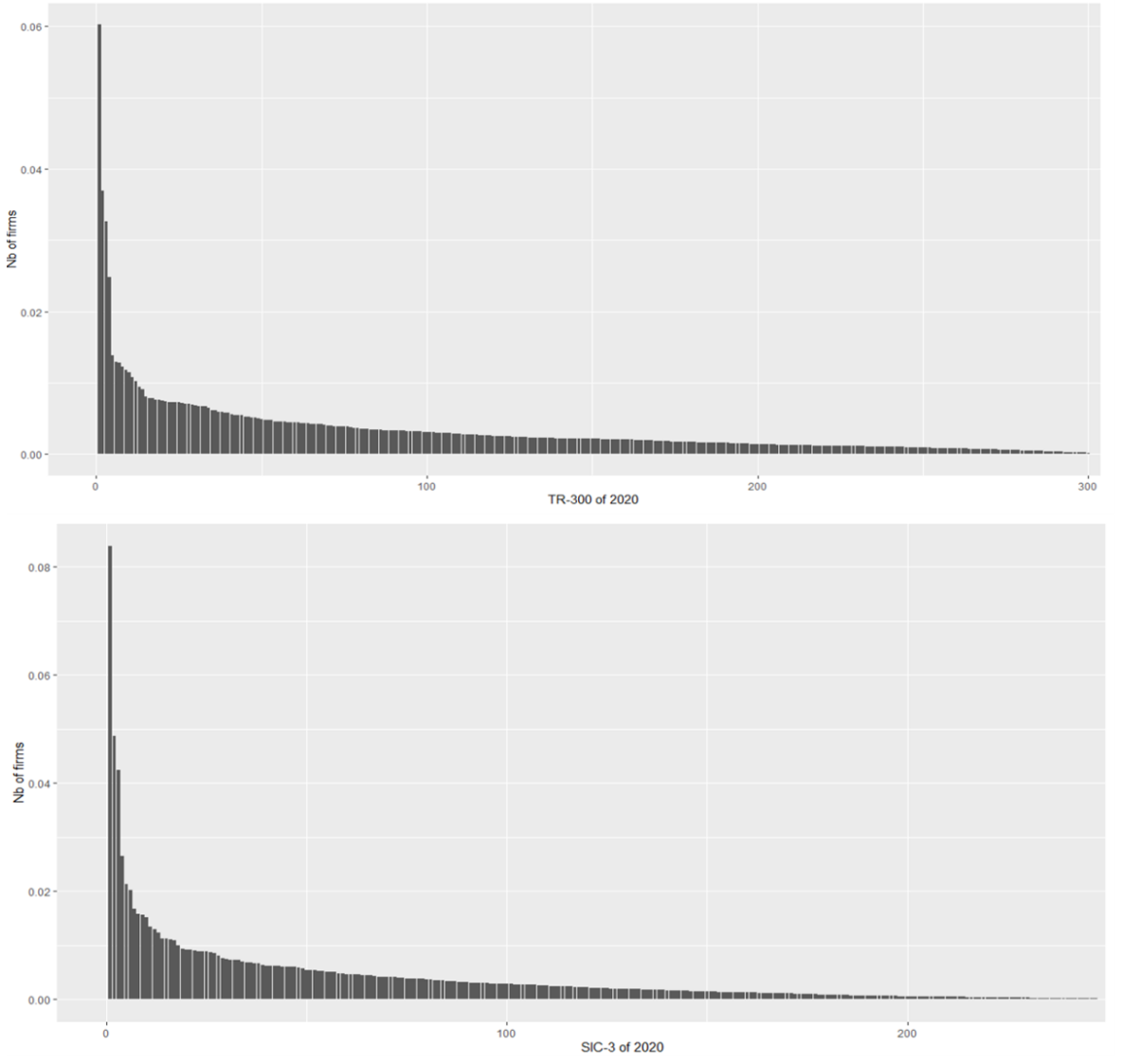


Figure 4: Distribution of the number of firms in each industry on the global market

However, this method has some drawbacks. First, the SIC codes I use to identify the single-segment firms are not historical data which can bias my results. As I mentioned before, a single-segment firm in the past can become a multi-segment firm later. Using the latest value will upward bias the proportion of the multi-segment firm in past years. Second, this method relies on the SIC codes to identify firms' segments, making it dependent on the SIC system. Is it possible to have a new industry classification that is totally independent of the existing industry classification? I will answer this question in the next section.

3.3.2 Network method

The network method is introduced by Hoberg and Phillips (2016). The main idea of this algorithm is to assign two firms as peers based on a threshold of business similarity. Assume there are N firms in the economy, and there will be (N^2-N)/2 permutations of unique pairs. Only a small fraction of firms will be membership pairs. SIC-3 generates 2% membership pairs. To mimic SIC-3, Hoberg and Phillips (2016) set a threshold of pairwise similarity to be 21.32% to generate the 2% membership pairs. To make the industries in this chapter comparable with SIC-3 and the TNIC, I choose a threshold that also generates 2% membership. The threshold found in Hoberg and Phillips (2016) is 21.32% while the ratio of this project is 21.52%. That's to say any two firms whose similarity is above or equal to 21.52% will be considered competitors under this technique.

Here I describe how to find out the threshold which generates 2% competitors. I first compute (N^2-N)/2 pairwise similarities and sort the similarities from highest to lowest. Then, I select the top 2% pairs. To describe the idea, assume there are 100 firms in the economy so that there will be 4950 pairs in total. The top 2% will be the first 99 firm pairs based on ranking the similarity score from high to low. I pick the similarity score of the 99th pairs as a threshold, say 0.25. Later, I select all firm pairs whose similarity scores are equal to or above this threshold. The number of firm pairs selected can be greater than 99 as 100th pairs may have the same similarity score as 99th pairs.

The number of words in each TRBD may bias the computations. For example, firm A with a long description can include many different terms, so it has positive similarities with many other firms. Suppose firm B has a short description with a few different terms, so it has a positive similarity with very few firms. The only difference between firms A and B could be simply the fact that one writes a more detailed description. To adjust for that, I will subtract the median similarity score for each firm. The median score is the median number of pairwise similarities between a given firm and the remaining of the firms in the economy. Suppose firm A's similarity with firm X is 0.8, this is the raw score. Now suppose firm A's median similarity with all other firms is 0.01. Then the adjusted similarity score for A and X is 0.8 - 0.01 = 0.79. I use the adjusted similarity score to compute the threshold that results in 2% membership pairs.

This method provides a given firm with all its competitors. One can know both the number of competitors for each firm as well as the pairwise similarities in the product space. Summing up all the similarity scores around a given firm shows the competitive

environment of this firm. A higher score means an intensive competition situation. Additionally, this method relaxes the transitivity of competitorship. When firm A and firm B are peers, and firm B and firm C are peers, firm A and firm C do not need to be peers if their business similarity score is below 21.52%. In reality, firm A and firm C may operate in different market segments or target different customer bases. Each firm might have a distinct product or service offering that caters to specific customer needs, resulting in minimal direct competition between them.

4 Empirical analysis

In this part, I will compare the text-based industry classification with the extant industry classifications and analyze the results.

4.1 Method and Data

4.1.1 Econometric method

Existing empirical work on the industry classification comparison focuses on the U.S. market, and I am the first one to explore this topic in the global environment. I use the method developed by Bhojraj et al. (2003) to compare the text-based method against the standard industry classifications. The idea is that good industry classification should generate industries with homogenous firms. One measure of homogeneity or similarity among different firms in terms of how the market values them is based on certain accounting metrics. Specifically, it focuses on valuation multiples assigned to key accounting measures such as earnings, book value of equity, and sales revenue. Valuation multiples are ratios that relate a company's stock price or market value to a specific accounting measure. Due to the data limitation, I only consider the Price-to-Book ratio in this study. The valuation multiples help investors and analysts assess the relative value of a company's stock and compare it with other companies in the same industry or sector. When there is a high degree of homogeneity in valuation multiples across firms, it suggests that the market perceives those companies to have similar levels of attractiveness, potential for growth, or risk. In other words, firms with similar accounting measures (e.g., similar earnings, book value of equity, or sales revenue) are being valued similarly by the market.

On the other hand, if there is significant variation in valuation multiples among firms with similar accounting measures, it indicates that the market perceives differences in their fundamental characteristics or prospects. This divergence might reflect

variations in factors such as growth potential, profitability, risk profile, market dominance, or other qualitative considerations.

By examining the extent of homogeneity or heterogeneity in valuation multiples, analysts can gain insights into how the market values different firms and assess the factors driving those valuations. It provides a comparative perspective on how the market perceives companies and can assist investors in making investment decisions or identifying potential opportunities or discrepancies in the market.

Financial researchers frequently have an interest in identifying companies that exhibit similar operating characteristics. This is done for purposes of comparison and control in their studies. To achieve this, I create industry portfolios using various industry classifications. I then examine how effectively the average industry ratio can explain important ratios at the individual company level. The financial ratios considered in their analysis include the return on assets (ROA), the return on equity (ROE), the Market-to-Book ratio, the Return on Net Operating Assets, the Assets Turnover, and the Leverage.

Variable name	**Full name**	**Definition**
ROA	Return on Assets	Net income divided by total assets
ROE	Return on equity	Net income divided by equity
Market_to_Book	Market to book ratio	The market cap divided by the balance sheet value of the equity
Price_to_Book	Price to book value	The share price divided by the book value per share.
RNOA	Return on net operating assets	The difference between Operating Income and Depreciation&Amortization divided by net sales
AssetTurnover	Assets turnover	Total Assets divided by Net Sales
Leverage	Leverage	Total Debt divided by Common Equity

Table 5: Variable definitions

Bhojraj et al. (2003) first build this method to compare the SIC, NAICS, GICS, and Fama-French industry classifications among S&P1500 firms between 1994 and 2001. Later work of Krazdil, Trottier and Zhang (2013) extends the research to a larger sample including all NYSE and NASDAQ firms from 1990 to 2009. Both papers suggest the GICS system outperforms other industry classification schemes in the U.S. market. When Kaustia and Rantala (2021) introduce the analysts' coverage approach, they also apply this econometric method to compare the efficiency of different industry classification systems.

The regression below shows the methodology of Bhojraj et al. (2003):

$$\text{VARIABLE}_{i,t} = \alpha_1 + \text{PEER\_AVERAGE}_{i,t} + \epsilon_{i,t}, \quad (1)$$

where VARIABLE $_{i,t}$ is the characteristic of interest observed for firm i in year t, and PEER_AVERAGE $_{i,t}$ refers to the average value among the competitors of firm i which excludes firm i itself. The adjusted R^2 demonstrates how much variation in the variable of firm i can be explained by its peer average. The higher the adjusted R^2, the more homogeneous the firms in one industry group are. Therefore, I mainly focus on the adjusted R^2 for the study.

4.1.2 Data description

To conduct the research, I download the historical SIC, NAICS, and GICS codes from Compustat for U.S. firms and Compustat Global for firms from other countries. Due to the limited data availability, I only collect NAICS-2 and NAICS-3 under the NAICS system. Table 6 shows the details of these industry classification systems used by my research. My main analyses are based on the same set of firm-year data with all the industry classification codes available. To achieve this end, I first compute the industry average for the network method because some firms can have no competitors under this measure. Later, I merge other rest industry classification codes data with these firms.

Name	Level	Title	Official Categories	Digits	Functional Categories
SIC	Tier 1	Major Division	11	1	10
	Tier 2	Major Group	83	2	69
	Tier 3	Industry Group	416	3	263
	Tier 4	Industry Sector	1,005	4	425
NAICS	Tier 1	Sector	20	2	25
	Tier 2	Subsector	99	3	110
GICS	Tier 1	Sector	11	2	11
	Tier 2	Industry Group	25	4	26
	Tier 3	Industry	74	6	72
	Tier 4	Sub-industry	163	8	171

Table 6: Official and functional categories

I obtain the financial variables from Datastream. All these variables are based on annual values at the end of the year. To carry out my analyses, I drop all firms with missing test variables and missing industry classification codes. To eliminate the outlier effect, I winsorize the test variables to 95% level.

To test the efficacy of different industry classification schemes, I first run all regressions on the global market. Later, I do the comparisons on the developed markets without the U.S.[20] as well as the developing markets. In the last stage, I test the industry classification systems on the U.S. market. Here, I also compare my global network method with the TNIC data built by Hoberg and Phillips (2016). The TNIC industry classification data can be downloaded from the internet.[21]

4.2 Main results

4.2.1 Global markets

Table 7 reports results for the regressions on the global markets. The network method provides a higher adjusted R^2 than competing classifications with only two exceptions (Assets Turnover and Leverage). For example, the adjusted R^2 of the network method of ROA is 17.8%, and the second best method only gets 13%. The variation in RNOA is the most difficult to clarify using industry membership, and the variations in ROA are the easiest. Improvements in explaining ROA range from 6.1% (against GICS-8) to 15.6% (against SIC-1), whereas the improvements in explaining RNOA range from 0 (against SIC-4) to 1.1% (against SIC-1). Besides the network method, the SIC system, especially SIC-4 best explains two variables which are RNOA and Assets Turnover. GICS-8 is the best to define the leverage ratio. To know how significant the relative enhancement from one method to the next is, I compute the average value of (R^2/highest R^2) for each method based on the results of Table 7. As shown in Figure 5, the network method obtains the highest value which is 0.93, and SIC4 and GICS-8 get 0.83 and 0.81 respectively. The weakest scheme gets a value of 0.26. The difference between the network method and the next best method is 10%. Based on these values, using the global network method provides an increase in explanatory power compared to other methods.

This effect is a novel finding in the literature as no previous studies have compared these systems on the global market. While countries can develop their own industry classification systems to capture the similarity among firms, a widely accepted standard is necessary for firms and academics to understand global market trends or compare firms from different countries. The text-based method presents a valuable opportunity to explore this domain and provides a more comprehensive and standardized approach

[20] They are the following countries and areas: Hong Kong, Sweden, Austria, New Zealand, Japan, South Korea, Netherlands, Australia, Switzerland, Taiwan, Portugal, Cyprus, France, Denmark, United Kingdom, Belgium, Italy, Spain, Norway, Canada, Luxembourg, Finland, Germany, Singapore, Ireland, Israel, South Korea.

[21] See the link: http://hobergphillips.tuck.dartmouth.edu/industryclass.htm.

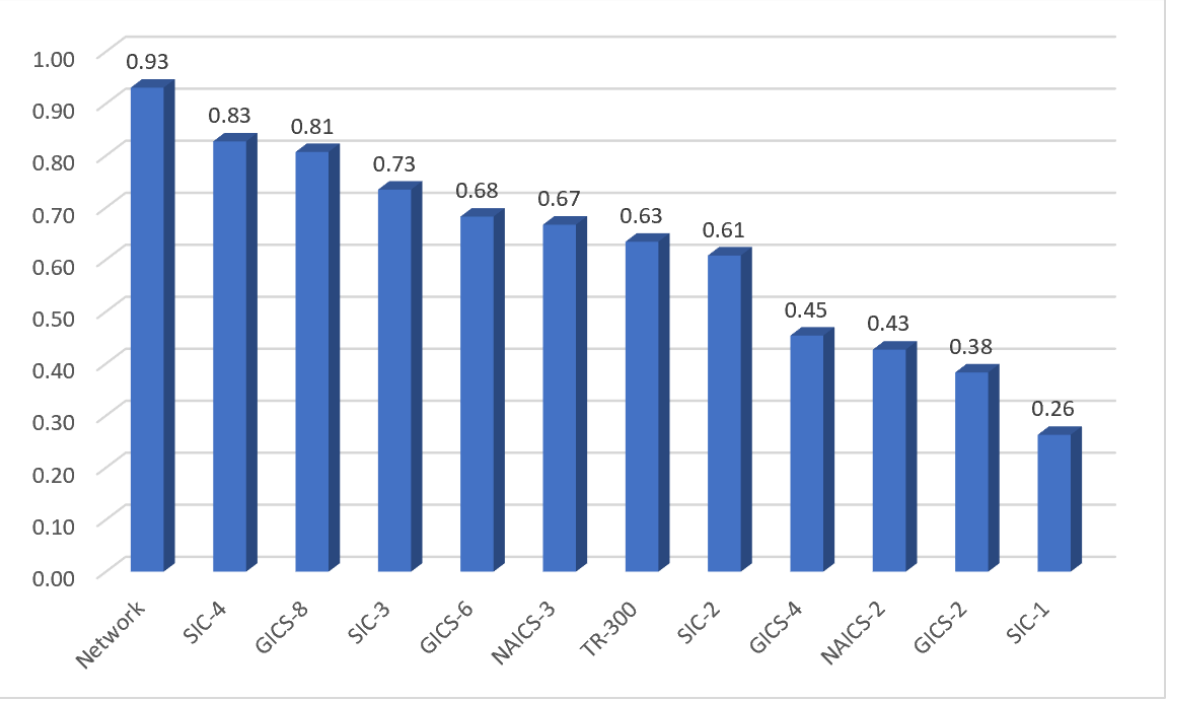


Figure 5: Average performance relative to best results on global markets

to analyzing global market trends.

4.2.2 Developed markets vs Developing markets

Table 8 shows the results of the developed countries excluding the U.S. The network method comes first place as it best explains 4 variables. The results are not very different from those of the global market except for the RNOA where the network method loses the power to better interpret this variable. The SIC family is still the second one and the only change is that SIC-3 becomes the champion to explain the Assets Turnover variable. Figure 6 shows the extent to which the network method outperforms other schemes. From this graph, one can observe that the difference between the network measure and the next is 0.2, which is smaller than that on the global markets. The reason can be either that the text-based method is weaker to define the industries on developed markets, or, more likely, the standard industry classification systems work much better on developed markets relative to developing ones.

Things become much more interesting on developing markets. The network method provides a higher adjusted R^2 than competing classifications with only one exception (Leverage). Table 9 shows the result of the developing countries, regarding RNOA, there is no big difference across different industry classifications. For the Leverage

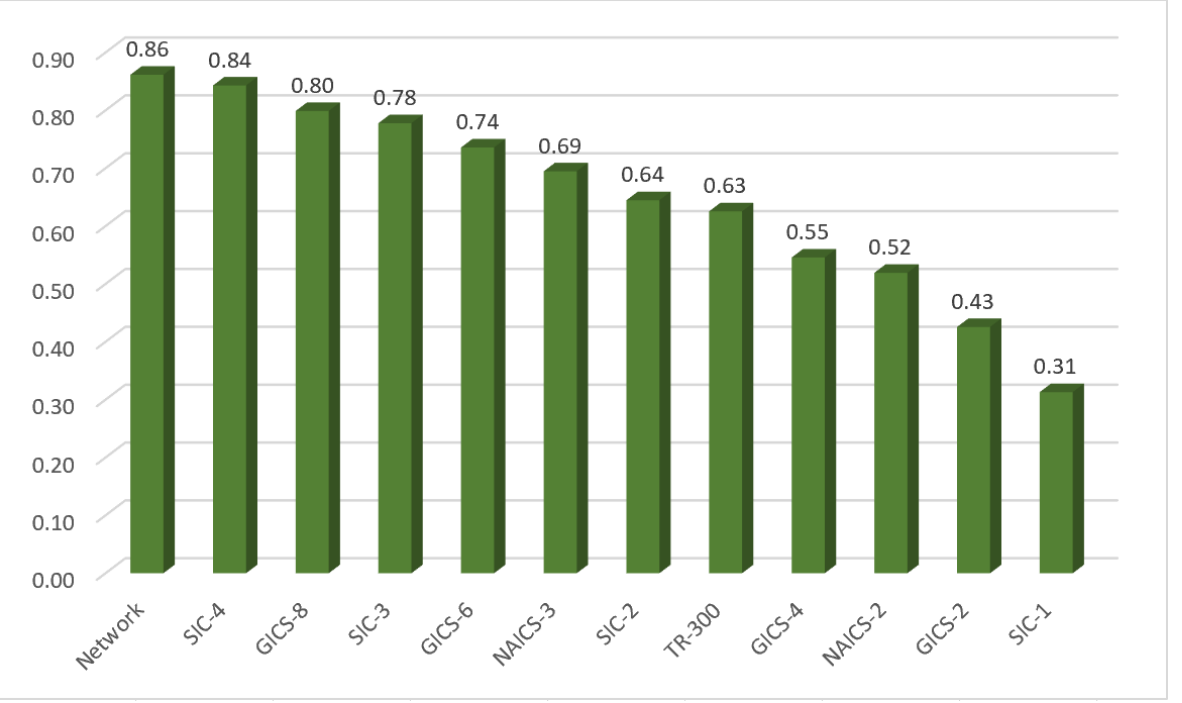


Figure 6: Average performance relative to best results on developed markets

variable, the network method is the second best to explain it and the difference between the network method and the best measure is 0.1%. Roughly speaking, one can trust the network method on all the test variables in terms of explanatory power. Besides, no other classification is comparable to the network method as GICS-8 or NAICS-3 only best explains one variable. Figure 7 shows the average performance relative to the best measure of developing markets. It shows this network method is by far the best approach. The gap between the network method and the second best is around 20%. From Figure 7, one can also observe that GICS-8 becomes the best measure among the traditional industry classifications. This is different from what is happening on developed markets. Another interesting finding comes from ROA: the adjusted R^2 of ROA is much smaller in developing countries compared to developed countries across all industry classifications.

4.2.3 The U.S. market

Table 10 shows the result for the U.S. market. The GICS group comes to first place in explaining 4 variables over 7 which are the Market-to-Book ratio, Price-to-Book ratio, Assets Turnover, and Leverage. The network method only best explains the ROA variable which achieves a value of 22.9%. Figure 8 shows the relative performance relative to GICS-8 on developing markets. One can observe that the network method

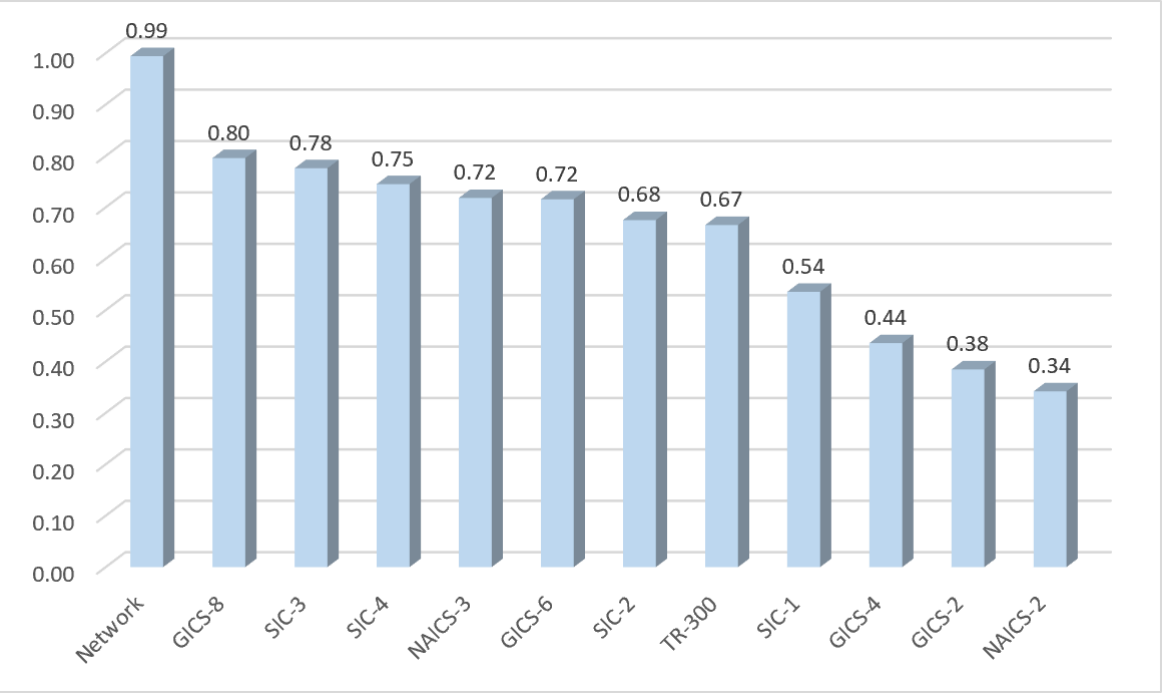


Figure 7: Average performance relative to best results on developing markets

becomes the fourth best. First, GICS-8 is the best approach in defining industries in the U.S. market is consistent with the results in the extant research such as Bhojraj et al. (2003) and Kaustia and Rantala (2021). Second, Kaustia and Rantala (2021) also show the TNIC method underperforms GICS-8 in their research. In contrast to the situation in developing markets, the adjusted R^2 of the ROA is the highest value across all industry classifications. When comparing all four regression tables together, one can find the network method best explains ROA across all countries, while GICS-8 best explains Leverage worldwide.

On the U.S. market, I also do a comparison between the network method based on TRBD with the TNIC. However, When I merge my dataset with 10-K firms, only around 71.14% of them can be matched.[22] Thus, I only do the tests with the ROA variable. The reason is if I test on all 7 variables, due to data availability, the sample size will drop significantly. Table 11 shows the result of the comparison between the network method based on TRBD with TNIC data based on the same firm-year level. From the results, no obvious difference is detected either for the adjusted R^2 or the coefficient of the variable. One can also observe, for a larger sample in the U.S. market, the adjusted R^2 of the network method improves from 22.9% to 30.6%.

[22]The low coverage may be due to the database disagreement problem because I merge TNIC, Compustat, and Datastream together. They are not fully compatible in terms of the identifier. For example, Datastream uses ISIN as the key for the merger while TNIC uses Compustat's gvkey.

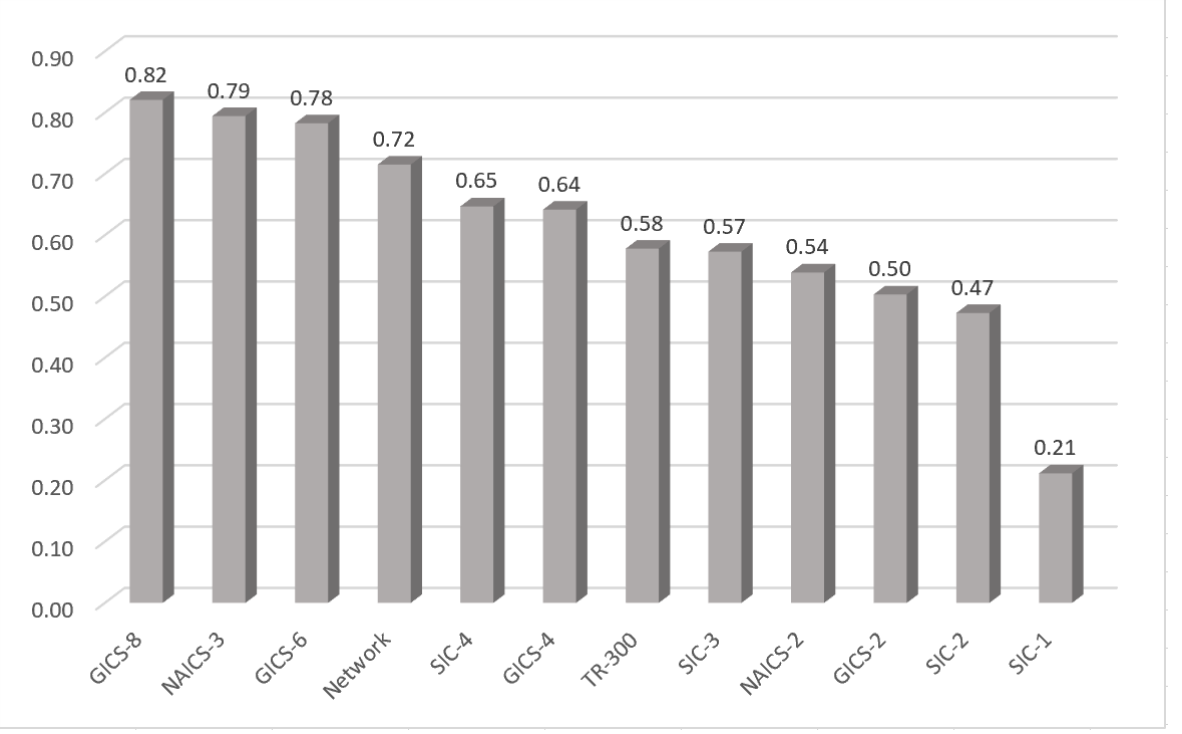


Figure 8: Average performance relative to best results in the U.S.

To conclude, the network method works best in all markets except the U.S. and it works especially well in developing countries.

4.2.4 Comments on the method and algorithm

Hoberg and Phillips (2016) show the fixed method which produces 300 industries outperforms the SIC and NAICS systems when comparing their efficacy. The fixed method (TR-300) in this project doesn't outperform all categories of SIC and NAICS. Two potential reasons can explain the inconsistency.

First, after forming 300 industries, Hoberg and Phillips (2016) add an additional step to improve the within-industry similarity by reassigning firms to alternative industries. Assuming 10,000 firms operate in the economy, after several iterations, 5,000 industries are generated and firm i best fits industry J. At the final stage, 300 industries are established, firm i may best fit industry Q which was not available when there were 5,000 industries. Hoberg and Phillips (2016) reassign firms to alternative industries until the within-industry similarity cannot be maximized. The assignment task requires massive computational power as the number of firms in this project is around nine times more than that of 10-K firms. Thus, I do not conduct this calculation, which can degrade the explanatory power of the fixed method.

Second, TRBD is much shorter than Item 1 of 10-K as I discussed before. This

	(TRBD) ROA	(TNIC) ROA
ROA_average	0.848***	0.855***
	(138.83)	(139.91)
N	43787	43787
adj. R^2	0.306	0.309

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: TRBD network method vs TNIC

requires more advanced technology to extract information from TRBD in order to keep as much useful information as possible. The bag-of-words method used in this chapter cannot detect the relatedness of similar texts such as "online" and "internet". This shortcoming matters more regarding short texts. There are other alternatives to the bag-of-words method, for instance, the word embeddings method. I will introduce two types of word embeddings method. i) Semantic fingerprinting, which is analogous to a human fingerprint, is meant to give a unique identity representation for any word. This method concentrates on the relationship between related concepts and each word is modeled on a sphere similar to a human's cerebral cortex (Numenta, 2011). Semantic fingerprinting is more powerful than the bag-of-words method to measure document similarity (Ibriyamova et al, 2016). For instance, when the business similarity between the text "commercial bank" and the text "financial investment" is 0 under the bag-of-words method, semantic fingerprinting technology results in 0.33 as a similarity score of the two texts. ii) Doc2vec, is a natural language processing tool for representing documents as a vector. The basic idea behind Doc2Vec is to train a neural network to predict a target word (or a next word) in the context of a given sequence of words in a document. In addition to learning the word embeddings, Doc2Vec also learns a unique vector representation, or "paragraph vector", for each document in the corpus. Empirical results show that Paragraph Vectors outperform bag-of-words models as well as other techniques for text representations.[23]

[23]See Le and Mikolov (2014) for more details on this technique.

4.3 Size effect

Bhojraj et al. (2003) have investigated the effect of firms on the comparison of different industry classification schemes. They divide firms into three groups based on their market size. Bhojraj et al. (2003) suggest the GICS system works best for big firms. Following their research, I divide the firms in the dataset into three groups based on their market value. The market value is from Datastream on an annual basis. Tables 12, 13, and 14 show the results of the large firms, mid-cap firms, and small firms. One can observe that the network method is the best to explain the ROA, ROE, Market-to-Book ratio, and Price-to-Book ratio for large firms and mid-cap firms. From Table 14, the network method is found to be best to interpret all rest of the variables for small firms. This finding is different from that of Bhojraj et al. (2003) that the existing industry classification usually works best for big-cap firms. In other words, the network method better defines small firms relative to big firms. To know the relative efficacy of the network method on big firms and mid-cap firms. I use the adjusted R^2 of the network method based on large-size, middle-size, and small-size firms to form Table 15. From this table, we can see the explanatory power for the mid-cap is the weakest. Figure 8 shows the average performance relative to the best results across sizes and it confirms this conclusion.

	ROA	ROE	Market_to _Book	Price_to_Book	RNOA	AssetTurnover	Leverage
Large	12.3%	5.1%	**18.1%**	**18.6%**	0.1%	0.7%	**6.3%**
Mid	19.0%	**6.8%**	11.9%	12.7%	0.6%	0.5%	3.6%
Small	**19.2%**	5.6%	7.9%	9.5%	**1.9%**	**1.9%**	2.7%

Table 15: Comparison of network method across firm sizes

4.4 Applications

4.4.1 An example

It is also interesting for academics to separate firms' foreign competitors from domestic ones when validating the industry classifications. The text-based method allows me to easily achieve this end. In this part, I only consider the ROA variable. Regression 2 shows the original regression which only takes the domestic competitors into account when forming the portfolios. Regression 3 adds one additional variable which is the foreign competitors. Regression 4 further decomposes foreign competitors into those from developed countries and developing countries. Here, I check the "b" coefficients,

and whether the adjusted R^2 improves when adding more variables. Table 16 shows the regression results. No obvious improvement of the adjusted R^2 is detected. However, all the coefficients are significant at a confidence level of 99%. One can observe the coefficient of the domestic average ROA decreases from 0.93 to 0.721 after adding the foreign average ROA variable. It suggests that these two variables are not fully independent of each other. When further decomposing the coefficient of the foreign average ROA, one can observe that the coefficient of the developed markets' average ROA is larger relative to the one from developing markets.

$$\text{ROA}_{i,t} = \alpha_1 + b_1 \text{ROAd\_peer}_{i,t} + \epsilon_i, \tag{2}$$

$$\text{ROA}_{i,t} = \alpha_1 + b_1 \text{ROAd\_peer}_{i,t} + b_2 \text{ROAf\_peer}_{i,t} + \epsilon_i, \tag{3}$$

$$\text{ROA}_{i,t} = \alpha_1 + b_1 \text{ROAd\_peer}_{i,t} + b_2 \text{ROAf\_dev\_peer}_{i,t} + b_3 \text{ROAf\_devping\_peer}_{i,t} + \epsilon_i, \tag{4}$$

	(Regression 2) ROA	(Regression 3) ROA	(Regression 4) ROA
ROA_avrg_domestic	0.930***	0.721***	0.730***
	(424.34)	(235.76)	(242.47)
ROA_avrg_foreign		0.438***	
		(96.74)	
ROA_avrg_developed			0.342***
			(85.70)
ROA_avrg_developing			0.178***
			(21.11)
N	505327	505327	505327
adj. R^2	0.263	0.276	0.276

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ROA_avrg_domestic is the average industry ROA based on the domestic peers.

ROA_avrg_foreign is the average industry ROA based on foreign competitors.

ROA_avrg_developed is the average industry ROA based on rivals from developed markets.

ROA_avrg_developing is the average industry ROA based on competitors from developing markets.

Table 16: Regression results of the new validation method

4.4.2 Further discussion

The network method can also be applied in many other research. I give two concrete examples here.

First, the global text-based network method can be deployed to measure foreign competition. Extant literature mainly uses the changes in tariffs to measure foreign competition. For example, Srinivasan (2020) shows firms are more likely to make horizontal acquisitions in response to increased foreign competition (reduction of tariffs), especially for financially unconstrained firms.[24] Alfaro, Conconi and Fadinger (2016) find evidence that output prices are a key factor of vertical integration and suggest higher tariffs will lead firms to conduct more vertical integration as product prices rise in the domestic market. This method has two limitations. First, it can only measure industries involved in tariff changes. With globalization creating new opportunities for economic growth and promoting social and political change, industries are facing more intensive foreign competition even when tariffs remain unchanged. The network method addresses this limitation by calculating foreign competition directly around any given firm, regardless of whether it resides in industries experiencing tariff changes or not. This provides a more comprehensive view of a firm's competitive position in the global market. Second, the tariff method provides a binary measure that does not provide information on the magnitude of the change. In contrast, the network method can quantify the magnitude of foreign competition faced by a firm by summing up all the foreign pairwise similarities of the firm. This information is crucial for researchers interested in understanding the relationship between competition and vertical integration. For example, Aghion et al. (2006) suggest a U-shaped relationship between competition and vertical integration. Using the network method, researchers can test whether foreign competition behaves in a similar way. In contrast, the tariff method is unable to provide this level of detail.

Second, the global network method provides a means of gauging the horizontal similarity in product space across industries and across countries. This is important for the study of competition as it enables firms to better understand their competitive position in the global market. When products are homogeneous across different countries, consumers in different markets are likely to choose products based on price and quality, rather than national origin or branding. This creates a highly competitive environment, where firms must compete on price, quality, and efficiency to succeed. By measuring product homogeneity, firms can identify their strengths and weaknesses relative to their competitors. Additionally, measuring product homogeneity across

[24] Also see Fresard (2010) and Fresard and Valta (2016).

countries can help firms identify opportunities for expansion into new markets. If a firm's product is highly homogeneous across different countries, it may be easier to enter new markets and compete with established players, as consumers are already familiar with the product and have similar preferences.

4.5 Conclusion

This study aims to introduce a novel approach to defining competitors on global markets. The standard industry classifications commonly used in the literature have some limitations which restrict the research on product competition. For instance, they cannot capture product homogeneity across countries and fail to update changes within an industry or the emergence of new industries on a timely basis. To address the limitations, a new dynamic text-based global industry classification system has been developed in this chapter. It offers a more flexible approach to industry classification and outperforms standard industry classifications in generating homogeneous groups, especially on developing markets. The outstanding performance of this method demonstrates its potential in addressing numerous important issues in today's global markets, where both globalization and deglobalization are taking place. With this method, one can easily determine the number of competitors and pairwise business similarities around any given firm. By summing up all the pairwise similarities, one can identify the competitive environment of given firms. My method provides a better alternative to measuring foreign competition relative to changes in tariffs as it tells us about the magnitude of foreign competition around a firm. Moreover, this new technique facilitates the calculation of foreign competition by allowing for easy separation of foreign competitors from domestic ones. Overall, this innovative methodology provides researchers with a more comprehensive tool to study industries in a rapidly changing economic landscape.

References

[1] Aghamolla, C. and Thakor, R.T., 2022. IPO peer effects. *Journal of Financial Economics*, 144(1), pp.206-226.

[2] Aghion, P., Griffith, R. and Howitt, P., 2006. Vertical integration and competition. *American Economic Review*, 96(2), pp.97-102.

[3] Albuquerque, R., Koskinen, Y. and Zhang, C., 2019. Corporate social responsibility and firm risk: Theory and empirical evidence. *Management Science*, 65(10), pp.4451-4469.

[4] Alfaro, L., Conconi, P., Fadinger, H. and Newman, A.F., 2016. Do prices determine vertical integration?. *The Review of Economic Studies*, 83(3), pp.855-888.

[5] Bereskin, F., Byun, S.K., Officer, M.S. and Oh, J.M., 2018. The effect of cultural similarity on mergers and acquisitions: Evidence from corporate social responsibility. *Journal of Financial and Quantitative Analysis*, 53(5), pp.1995-2039.

[6] Bernard, D., Blackburne, T. and Thornock, J., 2020. Information flows among rivals and corporate investment. *Journal of Financial Economics*, 136(3), pp.760-779.

[7] Bhojraj, S., Lee, C.M. and Oler, D.K., 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5), pp.745-774.

[8] Bloom, N., Schankerman, M. and Van Reenen, J., 2013. Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4), pp.1347-1393.

[9] Bonaime, A., Gulen, H. and Ion, M., 2018. Does policy uncertainty affect mergers and acquisitions? *Journal of Financial Economics*, 129(3), pp.531-558.

[10] Brown, S.J., Goetzmann, W., Ibbotson, R.G. and Ross, S.A., 1992. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4), pp.553-580.

[11] Bustamante, M.C. and Frésard, L., 2021. Does firm investment respond to peers' investment? *Management Science*, 67(8), pp.4703-4724.

[12] Campello, M. and Gao, J., 2017. Customer concentration and loan contract terms. *Journal of Financial Economics*, 123(1), pp.108-136.

[13] Chen, M.A., Wu, Q. and Yang, B., 2019. How valuable is FinTech innovation? *The Review of Financial Studies*, 32(5), pp.2062-2106.

[14] Eaton, G.W., Guo, F., Liu, T. and Officer, M.S., 2022. Peer selection and valuation in mergers and acquisitions. *Journal of financial economics*, 146(1), pp.230-255.

[15] Fama, E.F. and French, K.R., 1997. Industry costs of equity. *Journal of Financial Economics*, 43(2), pp.153-193.

[16] Florackis, C., Louca, C., Michaely, R. and Weber, M., 2023. Cybersecurity risk. *The Review of Financial Studies*, 36(1), pp.351-407.

[17] Frésard, L., 2010. Financial strength and product market behavior: The real effects of corporate cash holdings. *The Journal of Finance*, 65(3), pp.1097-1122.

[18] Frésard, L., Hoberg, G. and Phillips, G.M., 2020. Innovation activities and integration through vertical acquisitions. *The Review of Financial Studies*, 33(7), pp.2937-2976.

[19] Frésard, L. and Valta, P., 2016. How does corporate investment respond to increased entry threat? *The Review of Corporate Finance Studies*, 5(1), pp.1-35.

[20] Guenther, D.A. and Rosman, A.J., 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics*, 18(1), pp.115-128.

[21] Hoberg, G. and Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10), pp.3773-3811.

[22] Hoberg, G. and Phillips, G., 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), pp.1423-1465.

[23] Hoberg, G. and Phillips, G.M., 2022. Scope, Scale and Concentration: The 21st Century Firm (No. w30672). *National Bureau of Economic Research.*

[24] Hrazdil, K., Trottier, K. and Zhang, R., 2013. A comparison of industry classification schemes: A large sample study. *Economics Letters*, 118(1), pp.77-80.

[25] Ibriyamova, F., Kogan, S., Salganik-Shoshan, G. and Stolin, D., 2017. Using semantic fingerprinting in finance. *Applied Economics*, 49(28), pp.2719-2735.

[26] Ilhan, E., Krueger, P., Sautner, Z. and Starks, L.T., 2023. Climate risk disclosure and institutional investors. *The Review of Financial Studies*, 36(7), pp.2617-2650.

[27] Kaustia, M. and Rantala, V., 2021. Common analysts: method for defining peer firms. J*ournal of financial and quantitative analysis*, 56(5), pp.1505-1536.

[28] Krishnan, J. and Press, E., 2003. The north american industry classification system and its implications for accounting research. C*ontemporary Accounting Research*, 20(4), pp.685-717.

[29] Le, Q. and Mikolov, T., 2014, June. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[30] Li, K., Mai, F., Shen, R. and Yan, X., 2021. Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7), pp.3265-3315.

[31] Numenta. 2011. Hierarchical Temporal Memory Including HTM Cortical Learning Algorithms. *Technical Report*, Version 0.2.1

[32] Pearce, E., 1957. History of the Standard Industrial Classification, Prepared for the Executive Office of the President. *Bureau of the Budget*.

[33] Peters, R.H. and Taylor, L.A., 2017. Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2), pp.251-272.

[34] Ramnath, S., 2002. Investor and analyst reactions to earnings announcements of related firms: An empirical analysis. *Journal of Accounting Research*, 40(5), pp.1351-1376.

[35] Srinivasan, S., 2020. Foreign competition and acquisitions. *Journal of Corporate Finance*, 60, p.101484.

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage	obs
Network	**17.8%**	**6.0%**	**13.3%**	**14.1%**	**1.4%**	1.0%	3.0%	253406
TR-300	13.0%	3.7%	9.3%	9.9%	0.8%	0.8%	2.0%	253406
SIC-1	2.2%	1.5%	4.6%	4.8%	0.3%	0.3%	1.4%	253406
SIC-2	5.6%	3.3%	8.3%	8.7%	0.9%	0.8%	3.7%	253406
SIC-3	8.3%	3.7%	9.8%	10.2%	1.2%	1.0%	4.0%	253406
SIC-4	11.6%	4.4%	9.9%	10.3%	**1.4%**	**1.2%**	4.1%	253406
NAICS-2	4.3%	2.5%	5.9%	6.1%	0.7%	0.6%	2.0%	253406
NAICS-3	6.1%	3.4%	8.3%	8.8%	1.1%	0.9%	4.3%	253406
GICS-2	4.2%	2.3%	6.7%	6.9%	0.5%	0.5%	1.3%	253406
GICS-4	5.0%	2.1%	8.1%	8.4%	0.5%	0.5%	2.5%	253406
GICS-6	9.9%	3.3%	10.0%	10.4%	1.0%	0.7%	3.9%	253406
GICS-8	11.7%	4.5%	10.9%	11.5%	1.2%	0.9%	**4.4%**	253406

Table 7: R^2 Values from Peer Group Homogeneity Regressions: on global markets

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage	obs
Network	**17.5%**	**8.5%**	**14.6%**	**16.3%**	2.3%	1.5%	2.9%	117547
TR-300	13.0%	5.5%	10.1%	11.1%	1.7%	1.5%	1.7%	117547
SIC-1	4.5%	3.0%	5.4%	5.8%	0.7%	0.5%	1.8%	117547
SIC-2	9.3%	5.4%	9.8%	10.2%	1.6%	1.4%	3.9%	117547
SIC-3	11.6%	6.0%	11.6%	12.0%	2.5%	**1.8%**	4.2%	117547
SIC-4	13.2%	6.5%	11.6%	12.0%	**4.4%**	1.7%	3.9%	117547
NAICS-2	8.8%	5.2%	7.2%	7.8%	1.4%	1.0%	2.9%	117547
NAICS-3	10.1%	5.6%	9.6%	10.2%	2.0%	1.6%	4.3%	117547
GICS-2	5.7%	3.9%	7.9%	7.9%	1.1%	0.9%	1.8%	117547
GICS-4	6.7%	4.1%	10.3%	10.5%	1.2%	1.1%	3.1%	117547
GICS-6	11.1%	6.3%	12.5%	12.7%	1.9%	1.4%	4.0%	117547
GICS-8	12.9%	6.8%	13.2%	13.5%	2.4%	1.4%	**4.3%**	117547

Table 8: R^2 Values from Peer Group Homogeneity Regressions: on developed market excluding the U.S.

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage	obs
Network	**4.4%**	**2.8%**	**17.2%**	**17.4%**	**0.1%**	**0.4%**	2.3%	96500
TR-300	2.0%	1.6%	11.5%	11.7%	**0.1%**	0.3%	1.3%	96500
SIC-1	2.1%	1.6%	5.7%	5.7%	**0.1%**	0.3%	0.7%	96500
SIC-2	3.0%	2.4%	9.4%	9.7%	**0.1%**	0.1%	2.0%	96500
SIC-3	3.0%	2.5%	10.2%	10.4%	**0.1%**	0.3%	2.2%	96500
SIC-4	2.7%	2.3%	9.9%	10.1%	**0.1%**	0.3%	2.1%	96500
NAICS-2	2.2%	1.6%	6.7%	6.8%	0	0.1%	0.7%	96500
NAICS-3	3.2%	2.6%	9.6%	9.8%	**0.1%**	0.1%	**2.4%**	96500
GICS-2	2.4%	1.9%	7.6%	7.7%	0	0.1%	0.8%	96500
GICS-4	2.7%	2.1%	8.0%	8.2%	0	0.1%	1.2%	96500
GICS-6	3.5%	2.5%	10.2%	10.5%	**0.1%**	0.1%	2.1%	96500
GICS-8	3.5%	2.6%	11.5%	11.8%	**0.1%**	0.2%	**2.4%**	96500

Table 9: R^2 Values from Peer Group Homogeneity Regressions: on developing markets

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage	obs
Network	**22.9%**	2.0%	12.6%	13.0%	1.2%	1.4%	6.7%	39359
TR-300	16.5%	0.4%	9.3%	9.8%	1.0%	3.2%	4.5%	39359
SIC-1	2.2%	1.5%	4.6%	4.8%	0.3%	0.3%	1.4%	39359
SIC-2	5.6%	3.3%	8.3%	8.7%	0.9%	0.8%	3.7%	39359
SIC-3	8.3%	3.7%	9.8%	10.2%	1.2%	1.0%	4.0%	39359
SIC-4	11.6%	4.4%	9.9%	10.3%	**1.4%**	1.2%	4.1%	39359
NAICS-2	9.7%	0.5%	11.6%	11.9%	1.2%	1.1%	5.6%	39359
NAICS-3	13.0%	**4.5%**	14.0%	14.4%	1.0%	1.6%	10.2%	39359
GICS-2	12.0%	0.4%	11.6%	12.1%	0.9%	0.9%	4.8%	39359
GICS-4	16.7%	0.5%	14.5%	15.1%	0.9%	1.1%	7.9%	39359
GICS-6	19.2%	0.6%	**14.7%**	**15.4%**	1.1%	3.0%	9.9%	39359
GICS-8	19.0%	1.0%	14.1%	14.6%	1.1%	**3.9%**	**10.4%**	39359

Table 10: R^2 Values from Peer Group Homogeneity Regressions: in the U.S. market

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage
Network	**12.3%**	**5.1%**	**18.1%**	**18.6%**	0.1%	0.7%	6.3%
TR-300	8.6%	1.2%	12.3%	12.5%	1.2%	1.2%	3.4%
SIC-1	2.8%	0.2%	9.3%	9.3%	0	0	3.9%
SIC-2	5.2%	1.8%	12.4%	12.4%	0.2%	0.4%	7.5%
SIC-3	6.5%	2.4%	13.9%	13.9%	0.8%	1.0%	8.4%
SIC-4	8.8%	4.7%	13.9%	13.9%	**1.6%**	**1.8%**	8.5%
NAICS-2	3.7%	1.2%	9.8%	9.7%	0.1%	0	5.0%
NAICS-3	5.5%	3.4%	12.8%	12.7%	0.2%	0.2%	8.2%
GICS-2	4.3%	1.0%	10.5%	10.4%	0	0	4.9%
GICS-4	5.7%	1.3%	12.6%	12.6%	0.1%	0	6.6%
GICS-6	10.3%	1.8%	15.4%	15.6%	0.3%	0.1%	8.5%
GICS-8	11.0%	3.8%	16.5%	16.7%	1.2%	1.0%	**9.7%**

Table 12: R^2 Values from Peer Group Homogeneity Regressions: on large firms

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage
Network	**19.0%**	**6.8%**	**11.9%**	**12.7%**	0.6%	0.5%	3.6%
TR-300	15.0%	3.8%	8.2%	8.7%	0.4%	0.3%	2.6%
SIC-1	1.8%	0.1%	5.0%	5.2%	0.1%	0.1%	0.9%
SIC-2	5.3%	2.5%	7.9%	8.3%	0.4%	0.4%	5.4%
SIC-3	9.2%	3.4%	8.7%	9.0%	1.5%	1.0%	5.3%
SIC-4	12.9%	5.0%	8.7%	9.0%	**4.0%**	**4.5%**	5.1%
NAICS-2	3.7%	1.7%	6.2%	6.5%	0.3%	0.2%	2.8%
NAICS-3	6.0%	3.0%	8.0%	8.4%	0.3%	0.3%	**5.9%**
GICS-2	6.6%	0.2%	7.0%	7.3%	0.2%	0.2%	0.5%
GICS-4	8.8%	0.3%	8.6%	9.0%	0.3%	0.3%	4.2%
GICS-6	14.0%	0.6%	9.9%	10.1%	0.5%	0.4%	5.5%
GICS-8	14.4%	4.6%	10.0%	10.3%	0.7%	0.4%	5.6%

Table 13: R^2 Values from Peer Group Homogeneity Regressions: on mid-cap firms

	ROA	ROE	Market_to_Book	Price_to_Book	RNOA	AssetTurnover	Leverage
Network	**19.2%**	**5.6%**	**7.9%**	**9.5%**	**1.9%**	**1.9%**	**2.7%**
TR-300	16.9%	4.5%	5.3%	6.4%	1.0%	1.2%	2.2%
SIC-1	5.0%	3.5%	3.0%	3.4%	0.7%	0.5%	1.4%
SIC-2	12.0%	4.9%	6.1%	7.0%	1.4%	1.3%	2.5%
SIC-3	14.1%	4.6%	6.2%	7.3%	1.5%	1.5%	**2.7%**
SIC-4	15.8%	4.4%	5.8%	7.0%	1.6%	1.3%	2.4%
NAICS-2	10.0%	4.8%	4.4%	5.2%	1.1%	1.1%	2.3%
NAICS-3	12.4%	4.5%	5.9%	7.0%	1.5%	1.4%	2.6%
GICS-2	6.0%	3.8%	5.0%	5.2%	0.8%	1.1%	1.3%
GICS-4	6.4%	3.1%	5.7%	6.2%	0.8%	1.2%	1.7%
GICS-6	12.7%	4.9%	6.5%	7.5%	1.5%	1.5%	2.1%
GICS-8	16.3%	5.0%	7.1%	8.3%	1.7%	1.6%	2.5%

Table 14: R^2 Values from Peer Group Homogeneity Regressions: on small-cap firms

Chapter 2: Product relatedness and M&As in international markets: A text-based study*

Xiaojuan LIU†

March 2023

Abstract

I construct a measure of the pairwise relatedness of firms' products using textual analysis of their business descriptions to examine whether asset complementarity is a key factor in M&As on the global market. I find that high similarity between merging pairs' products can improve the profitability of acquirer firms, especially when acquirers operate in highly competitive markets. However, related mergers may not generate asset complementarity as no improvement in cost reduction, sale growth, or potential product differentiation is observed.

Keywords: M&As, asset complementarity, related mergers

*I am indebted to my Ph.D. supervisors Catherine Casamatta and David Stolin for their inspiration and guidance at every stage of my research work. I also thank Milo Bianchi, Sylvain Bourjade, Matthieu Bouvard, Patrick Coen, Laurent Fresard, Ulrich Hege, Sophie Moinas, Silvia Rossetto, Ph.D. Workshop in Finance at TSM and TSE, and especially Gerard Hoberg, for extremely valuable comments and suggestions on this thesis.

†University Toulouse 1 Capitole, xiaojuan.liu@tsm-education.fr

1 Introduction

The early work of Hart and Moore (1990) proposes that complementary assets should be combined under common ownership to mitigate the negative impact of incomplete contracts. Rhodes-Kropf and Robinson (2008) extend this perspective by developing a theory that implies assortative matching (i.e. like buys like). Utilizing a text-based method, Hoberg and Phillips (2010) suggest that similar buyers and sellers in the product space may generate asset complementarity, leading to the introduction of new products. Subsequent studies by Bena and Li (2014) and Lee et al. (2018) provide evidence that similar firms are likely to have asset complementarities in the realms of technology and human capital, respectively. However, these studies predominantly focus on the U.S. market, leaving the impact of asset complementarity on the global market underexplored.

This chapter examines the motivations and implications of asset complementarities in the global market. The primary implication is that when there are significant pairwise complementarities between firms' assets, value creation can result from mergers. Lee et al. (2018) suggest that the overlap in the acquirer and target firm workforce allows the acquirer to extract concessions from employees in the form of lower wages or retaining only the most productive components of the workforce. Fan and Goyal (2006) emphasize the importance of vertical relatedness of firms' industries for merger outcomes, while Bena and Li (2014) argue that technology synergies can arise from asset complementarities when merging pairs have technology similarities.

In this study, I use a sample of global M&A transactions from 2006 to 2018 to investigate the impact of asset complementarity. By focusing on publicly traded firms, an exploration can be conducted regarding the bargaining power of targets and the distribution of gains between firms that exhibit asset complementarities. The underlying concept revolves around the notion that merging with targets that share similarities in business can result in asset complementarity, ultimately leading to the introduction of new products. This effect becomes particularly pronounced when firms face intense competition in their respective markets. To illustrate, consider the scenario of a U.S. firm acquiring a French firm that operates in a similar product space. In this case, the U.S. firm can potentially benefit from complementary assets, such as leveraging the distribution channels of the French firm within France—a resource that the U.S. firm currently lacks. The presence of product similarity between the two firms allows the U.S. acquirer to gain valuable insights into the type of customers they can access and enables them to effectively manage the newly acquired distribution channels. Furthermore, if the product offered by the French target differs from those offered by the U.S.

acquirer's competitors, it presents an opportunity for the buyer to expand its range of products. The acquisition of the French firm can contribute to diversifying the buyer's product portfolio, potentially enhancing its competitive advantage in the market.

To better capture the business similarity between the buyer and seller, I use business descriptions of global public firms from Refinitiv to construct the main variables using a text-based methodology. The variable "PairSim" measures how similar the merging pairs are in the product space, while "GainProdDiff" measures the product distance from the seller to the buyer's close rivals. I also create variables to measure the competition faced by buyers and sellers which are "ProdSimAcq" and "ProdSimTarg" respectively. To investigate market reactions towards similar business mergers, I download transaction data from EIKON and financial data from Datastream to construct dependent variables. To test the market reaction, I conduct the event study over two windows from day -5 to day +5 and from day -10 to day +10 where day 0 is the announcement date. I calculate the cumulative abnormal return of the buyers and the combined entities over two event windows. I find that the market reacts positively to a transaction in which the buyer resides in a competitive industry while the target operates in a less competitive market. This finding is consistent with that of Hoberg and Phillips (2010).

To examine long-term performance, I construct profitability variables scaled by both total assets and sales, sales growth, and cost changes. The regression results show that profitability improves at the 1% level when buyers are in a competitive market. In addition, more value is created when the merging pairs share business similarities. Regarding the variable of "GainProdDiff", although it is positively significant in the event study, it becomes non-significant in the long-term performance analysis. Further, none of the main variables is found to be significant regarding the sales growth and cost changes. These results suggest that the value creation channel of business-similar merging pairs may not come from asset complementarity but other types of synergies on the global market.

Empirical research suggests that cross-border transactions differ from domestic ones from several perspectives. On the one hand, cross-border deals are perceived as riskier due to factors such as political uncertainty, exchange rate, and cultural difference (Ahern et al.,2015). On the other hand, cross-border deals also bring advantages such as the acquisition of technologies, foreign market entry, lower labor costs, etc. Does asset complementarity play a role in cross-border transactions? I do a separate analysis in this study to examine the cross-border transaction and the results suggest buyers experience an improvement in profitability if buyers facing intense competition and merging pairs are similar in the product space. However, the variable of "GainProd-

Diff" remains insignificant in both the event study and the long-term performance, which suggests the value creation of related cross-border mergers is not from the asset complementarity. Besides, I also verify domestic transactions beyond the U.S. market and draw a similar conclusion of the cross-border transactions.

When Hoberg and Phillips (2010) show how asset complementarity positively contributes to the combined entity based on U.S. transactions from the year 1997 to 2006. I restrict the sample to only U.S. firms to do a comparative study of Hoberg and Phillips (2010) from the year 2006 to 2018. The variables of interest are found to be significantly negative in this study and it suggests a harmful impact of business-similar mergers on buyers' long-term performance. The results are in contrast to that of Hoberg and Phillips (2010).

Research shows that the public status of the targets influences the outcomes of transactions. For example, researchers find that buyers experience significantly positive abnormal returns with private targets and non-significant abnormal returns with public sellers in the event study. This listing effect is detected in both the U.S. market (see Fuller, Netter, and Stegemoller 2002; Moeller, Schlingemann, and Stulz, 2004) and the European market (Faccio, McConnell, and Stolin, 2006). However, the reason behind the listing effect is unsolved. To investigate if the asset complementarity interacts with the public status of the targets, I further extend my sample to also include transactions with private targets. However, I find the buyers experience negative cumulative abnormal returns over the short event window when the buyers face intensive competition. This result is different from that of the public targets and it is the opposite of the listing effect. In the long run, buyers experience an improvement in profitability when buyers face more competition and the merging pairs are similar in the product space. "GainProdDiff" remains insignificant in this analysis. To conclude, the related mergers bring synergies to buyers but not through asset complementarity regardless of the public status of targets.

The difference between the findings in this study and that of Hoberg and Phillips (2010) may come from i) the study period. There is no overlap in the time period between the two studies. ii) financial data source. Hoberg and Phillips (2010) collect financial data from Compustat and I download the data from Datastream. Hoberg and Phillips (2010) winsorize the long-term performance variables to the 1% level and I winsorize the data to the 10% level. However, outliers are still spotted in my data sample. iii) text data used to build the business similarity measures. Hoberg and Phillips (2010) use Item 1 of 10-K filings and I use the short business descriptions from Refinitiv which is much shorter although more homogeneous.

By exploring the impact of asset complementarity on public transactions on the

global market, including cross-border deals, domestic deals, U.S. transactions, and transactions with private targets, this study provides valuable insights for practitioners, regulators, and investors, further enhancing our comprehension of the dynamics of mergers and acquisitions in both domestic and international markets. The rest of the chapter is organized as follows: Section 2 introduces the literature and hypotheses. Section 3 explains the textual methodology used in this study as well as the construction of the main variables. Section 4 shows the sample used in this study and also demonstrates some characteristics of the data. Section 5 conducts the empirical analysis and Section 6 draws the conclusion and suggests the future work.

2 Literature and Hypotheses

2.1 Literature

Asset complementarity in mergers and acquisitions (M&As) pertains to the synergies or benefits that emerge when two firms with complementary assets join forces. Complementary assets are those that, upon the combination, enhance the overall value or performance of the merged entity by creating economies of scale or scope, boosting operational efficiency, or providing competitive advantages. These assets can be tangible, such as physical resources or technologies, or intangible, including human capital, expertise, or brand reputation. Hart and Moore (1990) underscore the importance of complementary assets being controlled by a common owner in a world with incomplete contracts, providing new insight into the motivations for mergers. Rhodes-Kropf and Robinson (2008) expand on this argument by developing a theory that suggests assortative matchings (e.g., like buying like) and providing evidence that firms with high market-to-book ratios typically acquire companies with similar ratios. Hoberg and Phillips (2010) introduce a text-based methodology, demonstrating that asset complementarity is a crucial factor influencing the success of M&A transactions. When firms with complementary assets merge, they can capitalize on their combined strengths, leading to sales growth, enhanced profitability, and the introduction of new products. Examples of asset complementarity in M&As encompass vertical relatedness (Fan and Goyal, 2006), technological relatedness (Bena and Li, 2014), and human capital relatedness (Lee et al., 2017).

Hoberg and Phillips (2010) posit that business-similar merging pairs can generate potential asset complementarity, especially when buyers face intensive competition. They use business descriptions from Item 1 of 10-K filings of the U.S. companies to create variables that measure product similarity between merging pairs, demonstrating

that their methodology outperforms traditional industry classification systems such as SIC codes in defining peers (Hoberg and Phillips, 2016). Their findings reveal that transactions create more value when buyers face intense competition and sellers operate in less competitive industries. Additionally, they suggest that asset complementarity can lead to the introduction of new products. While their work offers valuable insights, they focus solely on U.S. firms, leaving a knowledge gap regarding other regions.

This project addresses that gap by examining the impact of similar mergers in international markets. Global market transactions include both domestic and cross-border deals, with cross-border M&A activities being less studied and often perceived as riskier due to trust, hierarchy, and individualism differences (Ahern et al., 2015). Additional risks arise from political and exchange rate factors, resulting in different motivations behind international bids compared to domestic M&As (Shimizu et al., 2004). Empirical research suggests that cross-border M&As generally integrate poorly relative to domestic transactions and experience poor post-merger performance (Adedeji and Ayoush, 2017). However, international bids offer advantages such as the acquisition of technologies from other countries by technology-intensive industries to strengthen their competitive position (Stiebale et al., 2011), foreign market entry, learning from foreign cultures, and value creation (Shimizuet al., 2004). Anand et al. (2005) argue that a target's multinational geographic scope can enhance the acquirer's ability to transfer and exploit knowledge. However, the role of asset complementarity in cross-border deals remains underexplored, and this chapter seeks to fill that knowledge gap.

Exploring domestic transactions outside the United States provides valuable insights into the dynamics of mergers and acquisitions beyond the U.S. market. Understanding if asset complementarity impacts these transactions adds depth to our knowledge of domestic mergers on a global scale. Domestic mergers can vary between the U.S. and other countries due to several factors, including legal frameworks, market structures, cultural differences, and regulatory environments. Here are a few key reasons for the differences: i) Legal and Regulatory Frameworks: Each country has its own set of laws and regulations governing mergers and acquisitions. The legal requirements, disclosure norms, antitrust regulations, and shareholder rights can differ significantly between the U.S. and other jurisdictions. These variations shape the processes and outcomes of domestic mergers. ii) Market Structures: Market structures can vary across countries. For instance, the level of industry concentration, competitiveness, and the presence of dominant players may differ. In the U.S., for example, certain industries may be characterized by a larger number of players and intense competition, while other countries may have more concentrated markets dominated by a few key firms. These differences influence the motivations, strategies, and outcomes

of domestic mergers. iii)Economic and Industry Factors: Economic conditions and industry-specific factors can vary across countries. Diverse economic cycles, growth rates, industry maturity, technological advancements, and market trends can influence the timing and nature of domestic mergers. For example, emerging markets may witness more consolidation activities driven by rapid growth and market expansion opportunities, whereas mature economies may focus on strategic acquisitions to enhance competitiveness. Thus, testing the domestic transactions outside the U.S. market helps answer the question of whether asset complementarity influences the deals in the same way that it does for U.S. transactions.

Understanding the impact of the public status of targets on the outcome of transactions is a crucial aspect of studying mergers and acquisitions. Existing literature suggests that when the target is private, public buyers tend to experience significant positive abnormal returns, while the abnormal returns for public sellers are not statistically significant. This phenomenon, known as the listing effect, has been observed in both the U.S. market (Fuller, Netter, and Stegemoller, 2002; Hansen and Lott, 1996; Moeller, Schlingemann, and Stulz, 2004) and the European market (Faccio, McConnell, and Stolin, 2006).In their research, Faccio, McConnell, and Stolin (2006) explored various potential factors contributing to the listing effect, such as cross-border deals and the introduction of a new blockholder. However, none of these factors were able to fully explain the observed listing effect. Therefore, this chapter aims to investigate whether asset complementarity plays a role in the listing effect. By examining the influence of asset complementarity on transactions with both public and private transactions, this study aims to shed light on the underlying mechanisms driving the listing effect.

2.2 Hypotheses

Building on the work of Hoberg and Phillips (2010), this study examines the combined impact of market competition and asset complementarity in mergers and acquisitions. When a firm faces intense competition, it may seek ways to improve profitability, including merging with other firms. One approach to enhancing profits is to introduce new products, as product differentiation enables firms to distinguish their offerings from competitors and appeal to various market segments (Hotelling, 1929). Hoberg and Phillips (2010) argue that similar mergers can generate asset complementarity, and using text-based methods, they demonstrate that asset complementarity assists the buyer in introducing new products to the market, ultimately improving the buyer's post-merger performance.

In this study, the first objective is to investigate whether mergers of business-similar

firms generate potential asset complementarities that lead to increased profits, sales growth, or cost savings for the buyer following the merger. Secondly, the study assesses whether asset complementarity aids the buyer in differentiating its products. By examining the interplay between market competition and asset complementarity, this research aims to provide insights into the benefits and strategic implications of pursuing similar mergers, especially in the context of an increasingly competitive global landscape. Thus, three hypotheses are developed in this study and displayed below.

Hypothesis 1 (H1): Merging pairs having similar products may benefit from asset complementarities and experience better ex-post long-term performance.

Hypothesis 2 (H2): The gains from similar mergers are increasing in the relatedness of the competitive situation faced by the buyers.

Hypothesis 3 (H3): More value is created if the target's business is distinct from that of the buyer's close rivals resulting in product differentiations.

3 Textual methodology and variables

To test the hypothesis, a new text database of global companies' descriptions is required. This section will introduce the database used in this study and provides insights into the main variables applied in the empirical analysis.

3.1 Introduction to text data

10-K filings are widely used by researchers, as used in previous studies such as Hoberg and Phillips (2010). However, since 10-K filings are only available for U.S. public firms, this chapter employs a new dataset, the short business descriptions from Refinitiv formerly Thomson Reuters (hereafter referred to as "TRBD". There are several differences between TRBD and descriptions from 10-K filings that are important to note. First, the business descriptions in TRBD are much shorter, with an average of around 122 characters, compared to the variable length of Item 1 in 10-K filings, which can range from one page to more than 10 pages. Hoberg and Phillips (2010) exclude firms whose business description contains less than 1,000 characters. Second, TRBD is updated more frequently than 10-K filings, as it depends on both financial reports and transactions in which the firms are involved, while 10-K filings are only updated annually. Third, TRBD is prepared by a third party rather than by firms themselves,

suggesting less heterogeneity in the descriptions. One crucial characteristic of TRBD is its extensive country coverage, including 112 countries, making it the only dataset with comprehensive historical business descriptions for global public firms that the author is aware of. Therefore, TRBD provides a valuable resource for analyzing product competition in a global context beyond the limitations of 10-K filings.

The TRBD data can be obtained directly from Eikon, which serves as a terminal provided by Refinitiv. However, it is important to note that while the business description data is up-to-date, historical data cannot be accessed through Eikon. Consequently, to address this limitation, the historical TRBD data for the period spanning 2006 to 2017 is acquired from Refinitiv. An essential consideration when working with this dataset is the potential presence of survivorship bias, which refers to the incomplete coverage of delisted firms. If the dataset fails to adequately capture information on firms that are no longer active, it may result in a significant loss of potential insights. To assess the coverage of delisted firms within this dataset, the global universe of equities from both Datastream and Factset for the years 2006 to 2017 is downloaded and compared with the historical TRBD. Datastream offers valuable information, including market value, public status, ISIN (International Securities Identification Number), and other relevant data points.

Based on empirical analysis, it is determined that the coverage of current public firms in the dataset is approximately 96% when considering market value. The coverage of delisted firms stands at around 92%. These figures provide insights into the extent of coverage for both active and delisted firms in the industry classification analysis. To ensure comprehensive coverage of global public firms beyond 2017, the TRBD is downloaded quarterly from Eikon. On average, there are 48,280 firms each year with available TRBD data. Furthermore, there has been a notable increasing trend in the number of firms included in the dataset from 2006 to 2021. The coverage based on market value across the years is 95%, indicating a high level of representation of firms within the dataset.

3.2 Business similarity

This study uses the bag-of-words method to analyze the text data in TRBD. This method involves creating a dictionary of unique words from the entire collection of documents being analyzed. Each document is then represented by a numerical vector, where each element of the vector corresponds to a word from the dictionary, and the value indicates the presence of that word in the document. The vector is binary as 1 indicates the existence of the word and 0 means the absence. After obtaining the

vectors of the documents, the cosine similarity formula is applied to compute the distance among the vectors.

To build the dictionary, the process starts by removing the last sentence of each TRBD containing the words "acquired" or "merged" as these sentences typically do not contain useful keywords for the analysis of a firm's business but rather details of the latest acquisition. Only nouns and proper nouns are kept, and all numbers, punctuation, and stop-words are removed from the remaining text.[1] All words are converted to lowercase, and plural forms are transformed to singular forms to reduce the length of the standard vector and improve computation accuracy. Next, words that appear in more than 25% of TRBDs and less than three documents are discarded, as they are considered less informative according to Hoberg and Phillips (2010). After this process, each TRBD only contains keywords, and these keywords are combined to form the dictionary. To vectorize a TRBD, a binary vector is created by assigning 1 to the positions of the common word sets between the TRBD and the dictionary. The binary vector is expected to be sparse given the short length of the TRBD.

Once the vectors of TRBDs are obtained, the cosine similarity formula is applied to compute the similarity between two vectors, allowing for the measurement of product similarity between firms. The formula of cosine similarity for two vectors $\mathbf{A}$ and $\mathbf{B}$ is as follows:

$$\text{similarity } = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are components of vector $\mathbf{A}$ and $\mathbf{B}$ respectively.

The calculation is easier than the formula shows because the vectors are binary. The denominator of the formula is the square of the number of keywords in vector $\mathbf{A}$ times the square of the number of keywords in vector $\mathbf{B}$. The numerator is the common words shared by vector $\mathbf{A}$ and vector $\mathbf{B}$. For instance, assume $\mathbf{A}$ is (0,1,0,1) and vector $\mathbf{B}$ is (0,1,0,0). The two vectors only have one common word which means the numerator is 1. The number of keywords in $\mathbf{A}$ is 2 and the number of keywords in $\mathbf{B}$ is 1. Thus, the cosine similarity is 0.71.

3.3 Key variables on product similarity

With knowledge of how to measure business similarity among firms, the main variables in this study are constructed. The transaction data is obtained from EIKON, and further details will be introduced in the following section. The transaction data in-

[1] I apply the 'StopWords_Generic.txt' developed by Loughran and McDonald, See the website for more details: https://sraf.nd.edu/textual-analysis/stopwords/.

cludes the business descriptions of the buyer and seller, which will be used to calculate the variables.

Product similarity to rivals (10 Nearest): This variable quantifies the level of competition in an industry, and its value ranges from 0 to 1. It is calculated based on the average pairwise similarity between a given firm and its ten closest competitors in the product space. Specifically, for a given firm j, I compute its pairwise similarity with all other firms in the industry and identify the ten most similar firms as its rivals. If the average of the ten pairwise similarities is high, then the firm is considered to operate in a highly competitive market.

Pair similarity: For a given merging pair, the variable estimates the similarity of their products. This variable is bounded within the range of 0 to 1, wherein higher values indicate greater similarity between the businesses of the two firms. Conversely, a value of 0 implies no relationship between the two firms' businesses. Specifically, if the firm pair's businesses have no overlap in their TRBD keywords, then the variable would be 0.

Gain in product differentiation: For a given merger pair, this variable is calculated as the disparity between the product distance of the target company to the ten closest competitors of the acquiring firm and the product distance of the acquiring firm to its own rivals. Product differentiation is determined as one minus product similarity (Hoberg and Phillips, 2010). This variable serves to indicate the level of dissimilarity between the product of the target company and those of the acquiring firm's competitors. Essentially, it quantifies the extent to which the acquiring firm can enhance its product differentiation from its rivals through the acquisition of the target company. It's important to note that the variable can be negative if the target company is more similar to the buyer's rivals than the buyer itself.

4 Data

4.1 Overview of M&As on global market

Transaction data are downloaded from Eikon SCREENER. I require the "Date Announced" to be a date between 2006 to 2018 and be completed by the end of 2018. To be consistent with previous related research, see Erel et al. (2012) for example, I only include "merger", "acquisition of assets", and "acquisition" in the

sample. I also exclude the buyer or seller if they are a government agency or in the financial industry. I further require the deal value to be greater than or equal to 1 million dollars. The transaction data covers the important information about each deal including Announcement Date, Year completed, Acquirer PermID,[2] Acquirer Nation, Acquirer Business Description, Target PermID, Target Nation, Target Business Description, Consideration Structure, Deal Value, Cross Border Deal Flag. A total of 90,499 deals were recorded, but after eliminating the deals with unavailable target or buyer nations, 87,512 transactions are retained. These transactions had a combined value of 25.24 trillion dollars. Of these, 27,634 were cross-border deals, accounting for 8.9 trillion dollars of the total value. The acquirer firms in this study hail from 178 countries, while the target firms represent 190 nations worldwide. The top 20 buyer and seller countries are the United States, United Kingdom, China, Canada, Australia, Japan, Hong Kong, South Korea, France, Sweden, Germany, Singapore, Spain, Italy, Brazil, India, Malaysia, Netherlands, Russia, and Norway.

Table A presents a cross-country matrix detailing the top 20 countries with the highest number of firm acquisitions. In this table, the number across the diagonal represents the domestic transactions. Table B shows the number of transactions and the deal value for the top 20 countries and the total deal value is expressed in millions of dollars. Combining the information from two tables, the U.S. leads in acquisition activities, with U.S. firms acquiring a total of 27,299 companies, including domestic transactions. The five countries with the highest merger and acquisition activity are the U.S. (27,299), the UK (8,990), China (7,304), Canada (6,650), and Australia (5,021). Focusing on cross-border deals, the most sought-after target nation is the U.S., with 5,759 firms acquired during this period. The other four popular target countries are the UK (3,132), Canada (1,476), China (1,404), and Australia(1,399). The U.S. also stands out as the largest acquirer, purchasing 5,415 firms from other countries in total. The other four countries purchasing the most are the UK (3,480), Canada(2,470), Australia(1,434), and Hong Kong (1,219). An interesting case is Hong Kong, where firms acquired 1,219 companies from other countries, while only 602 firms were bought by foreign investors during the same period. This trend may suggest that Hong Kong firms are more inclined to invest abroad, while the region is less attractive to foreign investors.

[2]Refinitiv's previous identifier of firms in M&A database is "SDC CUSIP" which is documented by other literature such as Gregoriou et al. (2021). After the upgrade of Refinitiv in 2021, the new identifier becomes "PermID". The identifier is used to download the corresponding financial data such as price, market value from Datastream.

	U.S.	UK	CHN	Canada	CAN	JPN	HK	KOR	FRA	SWE	DEU	SGP	ESP	ITA	BRA	IND	MYS	NLD	RUS	NOR
U.S.	23,120	834	120	750	261	41	45	39	198	76	259	42	162	114	76	66	13	127	15	44
UK	847	6210	20	96	165	7	24	7	137	81	223	16	131	82	28	14	7	138	14	63
China	120	30	6,622	36	42	16	165	6	21	5	30	28	6	18	4	0	8	17	0	3
Canada	1,317	144	21	4,418	94	2	21	4	27	15	44	2	21	8	34	3	1	21	4	10
Australia	283	143	16	74	3,989	8	21	5	12	8	24	29	11	3	16	7	15	11	2	6
Japan	199	56	6	12	31	3,444	11	17	9	5	19	28	6	10	12	7	13	13	2	1
Hong Kong	62	59	661	17	45	17	1,357	13	3	4	9	40	5	2	9	0	22	2	1	2
South Korea	66	18	20	12	5	12	5	1,770	5	0	7	3	2	2	0	7	2	0	4	1
France	144	85	8	20	12	3	2	7	1,085	15	67	8	75	45	21	13	1	35	0	6
Sweden	104	68	4	16	8	0	0	2	30	1,006	50	3	16	12	3	2	2	23	10	77
Germany	154	86	9	8	24	7	3	1	49	28	669	8	39	24	4	7	3	38	2	14
Singapore	53	57	101	4	92	36	54	17	5	4	11	643	4	2	2	24	65	11	0	4
Spain	60	33	4	6	7	1	2	0	36	6	22	1	921	26	17	3	1	8	1	3
Italy	40	43	4	5	10	1	0	0	30	4	15	2	17	931	9	3	0	17	5	1
Brazil	20	7	1	1	2	0	0	0	3	0	1	0	2	0	942	1	0	1	0	1
India	133	38	3	12	8	2	0	1	12	3	14	15	8	3	4	675	4	7	0	1
Malaysia	4	18	11	3	21	2	10	0	0	0	3	41	3	1	1	2	756	0	0	0
Netherlands	86	72	4	17	13	2	4	6	41	22	46	5	19	15	6	4	2	305	13	8
Russia	15	8	1	3	0	0	0	0	1	0	4	1	1	2	0	0	0	6	711	0
Norway	31	30	0	8	7	2	0	0	10	66	12	4	17	5	6	1	0	2	1	386

Table A: Number of Cross-border M&A Deals between Country Pairs

Country	All deals by acquirer nation		Cross-border deals by acquirer nation		All deals by target nation		Cross-border deals by target nation	
	Number	Value	Number	Value	Number	Value	Number	Value
United States	27,299	$11,842,070.67	5,415	$1,939,692.52	27,670	$12,714,824.34	5,759	$2,770,797.78
United Kingdom	8,990	$1,916,187.32	3,480	$1,154,320.69	8,681	$2,063,057.26	3,132	$1,305,563.28
China (Mainland	7,304	$1,394,985.61	1,063	$246,025.29	7,737	$1,363,468.45	1,404	$134,469.12
Canada	6,650	$1,326,311.37	2,470	$742,675.51	5,640	$1,038,855.25	1,476	$458,675.97
Australia	5,021	$686,473.09	1,434	$255,922.78	4,989	$775,355.25	1,399	$359,487.82
Japan	4,002	$792,680.02	679	$428,565.74	3,627	$470,086.92	306	$106,159.64
Hong Kong	2,503	$515,242.93	1,219	$243,859.63	1,823	$319,668.93	602	$126,528.88
South Korea	1,993	$301,106.23	276	$61,143.61	1,910	$275,992.49	194	$36,424.86
France	1,816	$818,963.25	887	$537,269.65	1,881	$557,636.40	955	$276,021.06
Sweden	1,669	$264,766.23	779	$132,658.65	1,468	$270,682.11	573	$135,158.80
Germany	1,379	$616,016.97	860	$455,065.72	1,821	$601,571.08	1,305	$443,475.55
Singapore	1,349	$209,070.63	789	$134,945.25	990	$164,232.83	446	$90,945.66
Spain	1,286	$368,858.91	508	$210,083.12	1,581	$375,345.85	806	$216,705.78
Italy	1,203	$350,717.18	430	$119,297.01	1,400	$414,869.97	625	$186,214.89
Brazil	1,033	$328,807.70	292	$71,942.58	1,260	$381,703.04	519	$126,216.01
India	1,020	$179,904.18	447	$88,031.96	866	$151,591.28	294	$59,776.95
Malaysia	918	$93,119.69	222	$22,393.60	946	$93,942.60	247	$23,233.21
Netherlands	861	$414,588.26	615	$294,169.10	902	$417,193.86	668	$317,203.85
Russia	805	$166,293.75	183	$55,989.21	858	$151,131.72	234	$39,130.10
Norway	665	$144,591.77	322	$74,485.63	697	$172,941.92	349	$102,588.84

Table B: Descriptive statistics of number of M&A deals for each country

4.2 Financial data and variables

To investigate the market reaction towards transactions and the long-term performance, I require both the buyer and seller to be publicly traded during the study period (I include the private targets in later analysis), and this results in a total of 4,691 deals.

The financial data of corresponding deals are collected from Datastream. Cumulative Abnormal returns (CARs) are calculated by subtracting the return of MSCI World index[3] from the firm's daily return (see Ahern et al.,2015). Investors often need more time to respond to M&A deals on the global market, thus, I establish two event windows for the market reaction analysis. The first event window spans from day -5 to

[3]MSCI World is a Market Index based on the global market starting from 12/31/1969 with the mnemonic of MSWRLD$ in Datastream.

day 5, with day 0 representing the announcement date. The second event window extends from day -10 to day 10. The combined CAR is the average of the acquirer's and target's CAR, weighted by each firm's ex-ante market value (see Ahern et al.,2015). Following the requirement for all announcement return data to be available for transactions, the sample size is reduced to 3,017 deals. The variables are defined as follows.

ACAR1 and ACAR2: Buyers' abnormal return over the period of day -5 to day 5 and day -10 to day 10.

CCAR1 and CCAR2: The combined CAR is the average of the acquirer's and target's CAR, weighted by each firm's ex-ante market value over the period of day -5 to day 5 and day -10 to day 10.

To investigate the long-term performance of firms after M&A transactions, I focus exclusively on the buyer firms. Using Datastream, I download key financial metrics such as Earnings Before Interest and Taxes (EBIT), Total assets (TA), Net sales, and Cost of Goods Sold (Cost) for the years t+1, t+2, and t+4 (where t represents the year of completion). Next, I calculate the 1-year and 3-year changes in EBIT over TA, EBIT over Net sales, sales growth, and cost changes. To ensure the validity of the calculations, I only include observations where both Total assets and Net sales were non-zero. To benchmark the industry averages for these variables, I use Standard and Poors Global, which is denoted by the mnemonic LSBPGGL£ in Datastream and covers a total of 1,922 global public firms starting from 2006. For each firm in the Standard and Poors Global dataset, I collect the EBIT, TA, Net sales, Cost, as well as the SIC-3 code, based on which I calculate the industry average values. Finally, to obtain industry-adjusted financial variables for all buyers in the sample, I subtract the industry average values from the firm-level values.

Δ Profitability scaled by assets (PA1 and PA3): For a given firm j in year t, this variable is the change of operating income scaled by total assets from year t+1 to year t+2 or t+4, adjusted by industry average.

Δ Profitability scaled by sales (PS1 and PS3): For a given firm j in year t, this variable captures the change of operating income divided by net sale from year t+1 to year t+2 or t+4, adjusted by industry average..

Sales Growth (SG1 and SG3): For a given firm j in year t, this variable cap-

tures the change of net sale from year t+1 to year t+2 or t+4, adjusted by industry average.

Cost Changes (CR1 and CR3): For a given firm j in year t, this variable captures the change of cost from year t+1 to year t+2 or t+4, adjusted by industry average.

Table C presents the descriptive statistics of the variables. It can be observed that, on average, the pairwise similarity between merging pairs is approximately 0.25, while the product differentiation that the target company can provide to the buyer is also 0.25. The competitive environment for the target company is slightly more intense, with a value of 0.48, compared to 0.47 for the acquiring company. HHI in Table C is Herfindahl–Hirschman index and the computation of this variable is shown later in this Section.

In the short event window, the average cumulative abnormal return of the buyer is 0.01, while that of the combined entity is 0.015. For the longer event window, the average cumulative abnormal return of the buyer is 0.022, and for the combined entity, it is 0.023. Hoberg and Phillips (2010) find that the average abnormal return of buyers on the event day is 0, while that of the combined entity is 0.004. In the study by Ahern et al. (2015), the average cumulative abnormal return of the buyer over a window of day -1 to day +1 was 0.002, while that of the combined entity was 0.036. These findings suggest that the average cumulative abnormal returns of buyers and combined entities in this study are consistent with previous research. Any slight differences observed may be attributed to variations in the length of the event window.

Regarding the long-term performance, I get negative average sales growth of -0.124 over one year window and negative average profitability scaled by sales of -0.395 over one year window. In contrast, Hoberg and Phillips (2010) reported corresponding values of 0.035 and -0.004, respectively. The averages of other long-term variables, such as changes in cost, are positive and range from 0.003 to 0.338 in this study. However, Hoberg and Phillips (2010) find negative values ranging from -0.021 to -0.005 for the same variables. The differences observed between the averages of long-term variables in this study and Hoberg and Phillips (2010) can be attributed to two factors. Firstly, the study periods and sample countries differ. Hoberg and Phillips (2010) focus on U.S. public transactions from 1997 to 2006, while this project investigates global public transactions from 2006 to 2018. Secondly, the data sources used also vary. Hoberg and Phillips (2010) obtain financial data from Compustat, whereas the data for this study are sourced from Datastream.

Variable	Mean	Std Dev.	Min	Max	25th percentiles	75th percentiles	Obs.
		Panle A: Firm variables					
Product similarity to rivals (Buyer)	0.470	0.137	0.178	1.000	0.369	0.553	3017
Product similarity to rivals (Seller)	0.484	0.163	0.167	1.000	0.358	0.577	3017
Log total assets	7.410	2.344	0.131	13.264	5.813	9.056	3017
HHI based on SIC-3	0.222	0.246	0.027	1.701	0.075	0.231	3017
		Panel B: Transaction level variables					
ACAR1	0.010	0.175	-0.720	4.866	-0.049	0.047	3017
ACAR2	0.022	0.310	-0.791	8.924	-0.058	0.064	3017
CCAR1	0.015	0.118	-0.753	2.293	-0.029	0.037	3017
CCAR2	0.023	0.167	-0.974	3.568	-0.038	0.052	3017
Gain in product differentiation	0.252	0.140	-0.346	1.000	0.166	0.338	3017
Merging pair similarity	0.254	0.207	0.000	1.000	0.099	0.378	3017
		Panel C: Acquirer ex post performance					
1-Year Sale growth	-0.124	2.119	-32.609	5.721	-0.158	0.085	1221
3-Year Sale growth	0.159	1.750	-4.084	26.305	-0.349	0.185	1221
1-Year Cost change	0.004	0.581	-3.960	7.298	-0.198	0.099	1221
3-Year Cost change	0.199	4.649	-3.572	134.872	-0.417	0.154	1221
1-Year profitability (scaled by assets)	0.003	0.548	-4.554	2.403	-0.037	0.042	1221
3-Year profitability (scaled by assets)	0.085	0.500	-4.557	3.704	-0.037	0.081	1221
1-Year profitability (scaled by sales)	-0.395	13.025	-448.270	24.731	-0.074	0.088	1221
3-Year profitability (scaled by sales)	0.338	1.760	-5.313	24.613	-0.066	0.184	1221

Table C: Summary statistics of transactions

4.3 Pair Similarity of merging pairs and random pairs

According to the hypothesis, merging firms tend to be more similar in their business, which can lead to reduced information asymmetry and increased asset complementarity. To test this hypothesis, I focused on the year 2007, which had the highest number of deals (516) during the sample period. Specifically, I compared the business similarity of merging pairs in 2007 to that of randomly selected non-merging pairs in the same year. The non-merging pairs are selected from the global public firms with available business descriptions in the corresponding year.

Figure 1 depicts the distribution of business similarity values among the merging pairs, with the vertical axis representing the proportion of merging pairs and the horizontal axis representing the business similarity value (in percentage). The distribution ranges from 0 to 1, with the majority of values falling between 0.1 and 0.5. In contrast, Figure 2 shows the distribution of business similarity values among the randomly selected non-merging pairs, which ranged from 0 to 0.15. Interestingly, around half of these pairs were not similar in the product space at all, compared to less than 20% of merging pairs. These findings suggest that merging pairs are indeed more similar in the product space than non-merging pairs, which supports the hypothesis that business

similarity can play a role in M&A activity.

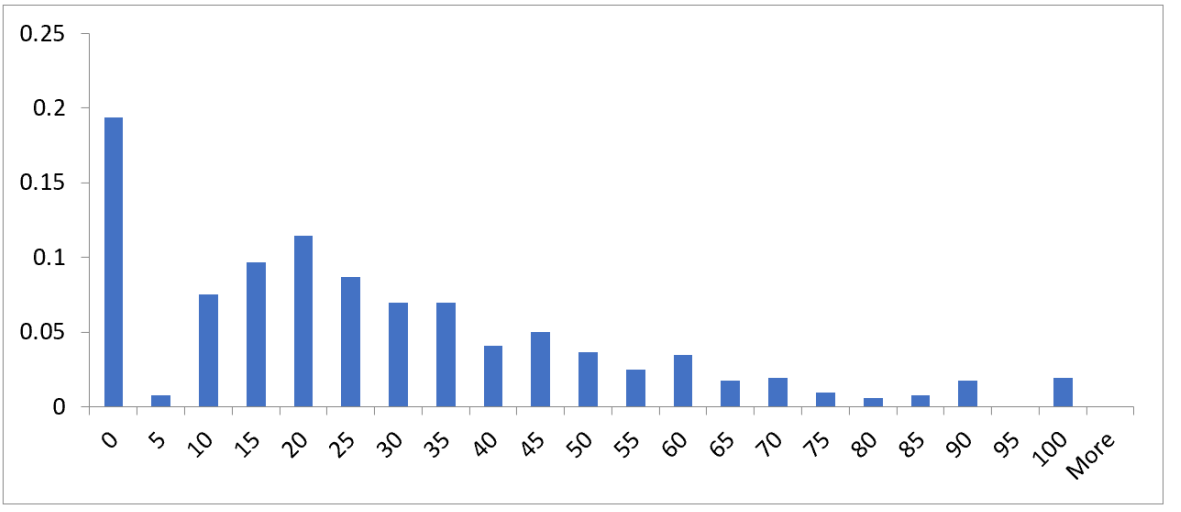


Figure 1: Distribution of product similarity for merger pairings

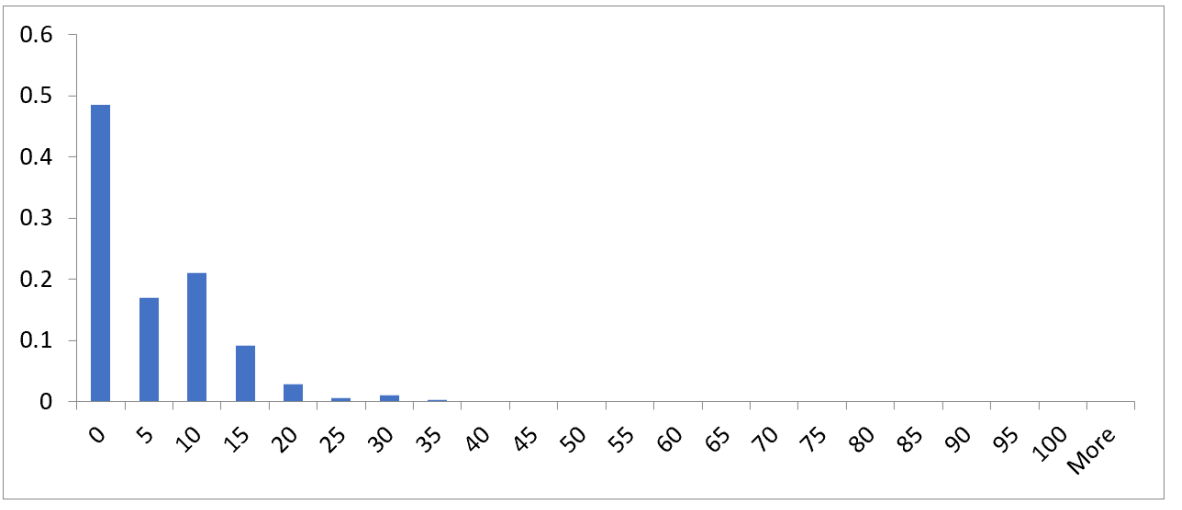


Figure 2: Distribution of product similarity for random firm pairings

Table D presents data on merger pairs during the given period that exhibit high product similarity (>=90 percentile) despite belonging to different SIC-3 codes.

4.4 Other control variables

I include firm-level control variables to capture the asset complementarity effect better.

Log total asset: The natural logarithm of the sum of the pre-announcement market value of the target and the acquirer in a transaction.

M&A Deal Identifier	Acquirer	Target	Acquirer SIC-3	Target SIC-3	Product similarity
34391274875	NBC Capital Cor	Suncoast Bancorp Inc	602	671	1
34391707370	Piper Capital Inc	Garson Resources Ltd	109	149	1
34391278497	WPS Resources Corp	Peoples Energy Corp	493	492	1
34391704692	InvestinMedia PLC	Avesco PLC	679	489	0.85
34391692480	Bowleven PLC	FirstAfrica Oil PLC	138	131	0.82
154085220066	Enlink Midstream LLC	EnLink Midstream Partners LP	131	492	0.81
34391328914	Kindred Healthcare Inc	RehabCare Group Inc	805	806	0.74
34391285666	Universal Compression Holdings Inc	Hanover Compressor Co	735	138	0.73
34393056067	Energy Xxi(Bermuda)Ltd	EPL Oil & Gas Inc	138	131	0.73
154083084275	Petrocapita Income Trust	Palliser Oil & Gas Corp	138	131	0.71
34391278303	UnionBancorp Inc	Centrue Financial Corp	602	603	0.71
154083011043	Sunko Ink Co Ltd	Kuo Ching Chemical Co Ltd	282	289	0.71
154082633901	Sunshine Bancorp Inc	Community Southern Holdings Inc	603	602	0.71
34392737140	Pine Cliff Energy Ltd	Geomark Exploration Ltd	131	104	0.71
34391288953	Harleysville National Corp	East Penn Financial Corp	602	671	0.71
34391276061	Alliance Financial Corp	Bridge Street Financial Inc	602	603	0.71
34391284223	Citizens & Northern Corp	Citizens Bancorp Inc	602	671	0.71
34391278543	National Bancshares Inc	Metrocorp Inc	602	671	0.71
34391723620	Aurium Resources Ltd	Haven Resources Ltd	104	109	0.67
34391801063	Blue Note Mining Inc	X-Ore Resources Inc	104	109	0.67

Table D: Merging firms with high business similarity (>95 percentile) but different SIC-3

Sales Herfindahl-Hirschman Index (HHI): This variable is a commonly accepted measure of market concentration that is computed by squaring the market share of each firm competing in a market and then summing the resulting numbers. I calculate HHI based on SIC-3 from Standard & Poors Global.

Same SIC-3 dummy: For a given merger pair, this variable indicates if the buyer and target reside in the same SIC-3 industry.

Cash only: For a given transaction, the variable is set to be one if the deal is paid only by cash.

Stock only: For a given transaction, this variable is set to be 1 if the deal is settled only by stock.

Cross border dummy: For a given transaction, this variable is set to be 1 when

the target company in the deal is not domiciled in the same country as the acquirer.

Considering cross-border deals are also included in this study, following Erel et al. (2012) and Ahern et al. (2015), I control for country-level variables. The country-level data are GDP Growth, GDP per capita, GDP per capita growth, and trade openness. The data is downloaded from the World Bank. Trade openness is calculated by summing up the total imports and exports divided by the GDP of a given country. In regressions, these variables are represented in the difference between the buyer's and seller's countries.

(GDP Growth)_b-s: The difference in the growth rate of gross domestic product between the buyer's and seller's countries.

(GDP per capita Growth)_b-s: The difference in the growth rate of gross domestic product (in USD) divided by the average population between the buyer's and seller's countries.

(Log GDP per capita)_b-s: The difference in the natural logarithm of gross domestic product (in USD) divided by the average population between the buyer's and seller's countries.

(Trade Openness)_b-s: The difference in the percentage of total imports and exports value to GDP between the buyer's and seller's countries.

5 Empirical Analysis

This section introduces the empirical analysis of the impact of asset complementarity on M&As.

5.1 Announcement return

First, I conduct an event study to test the impact of asset complementarity on announcement returns of buyers using regressions 1 and 2. In regression 1, I examine the relationship between the cumulative abnormal return of the buyers (ACAR) and the "ProdSimAcq" variable. The "ProdSimAcq" represents the level of product similarity of the buyer, which is determined by calculating the average pairwise similarity

between the buyer and its ten closest rivals in the global market. $\mathbf{X}$ is the set of control variables, t indicates the corresponding year while i refers to the corresponding deal. "Year F.E." is the year-level fixed effect and "Country F.E." means the country-level fixed effect. Regression 2 tests whether buyers experience higher announcement returns in the related mergers, especially when targets are different from buyers' rivals in the product space. "GainProdDiff" measures the product distance between targets and buyers' ten closest rivals in business and "PairSim" indicates the pairwise product similarity for a merging pair. Regression 3 examines the same relationship as regression 1 but for the combined entity of a transaction. In this equation, the variable "ProdSimTarg" represents the product similarity of the seller. Specifically, it quantifies the average pairwise similarity between the target company and its ten closest rivals in the global market. Similarly, regression 4 explores the same narrative as regression 2 but for the combined entity.

$$\text{ACAR}_{i,t} = \beta_0 + \beta_1 \text{ProdSimAcq}_{i,t} + \beta_2 \mathbf{X}_{acq,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \quad (1)$$

$$\text{ACAR}_{i,t} = \beta_0 + \beta_1 \text{GainProdDiff}_{i,t} + \beta_2 \text{PairSim}_{i,t} + \beta_3 \mathbf{X}_{i,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \quad (2)$$

$$\text{CCAR}_{i,t} = \beta_0 + \beta_1 \text{ProdSimAcq}_{i,t} + \beta_2 \text{ProdSimTarg}_{i,t} + \beta_3 \mathbf{X}_{i,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \quad (3)$$

$$\text{CCAR}_{i,t} = \beta_0 + \beta_1 \text{GainProdDiff}_{i,t} + \beta_2 \text{PairSim}_{i,t} + \beta_3 \mathbf{X}_{i,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \quad (4)$$

My hypothesis is that when a target firm can provide complementary assets to the buyer, it will elicit a positive reaction. Thus, the coefficients of "GainProdDiff" and "PairSim" are supposed to be positively significant. Table 1 presents the regression results of abnormal returns of buyers during two event windows. The coefficient of "GainProdDiff" is notably positive and statistically significant at the 1% level with a t-value of 2.73 during the event window from day -5 to day 5, and significantly positive at the 5% level with a t-value of 2.18 for an extended event window.

Furthermore, the combined abnormal return regression results demonstrate that when the buyer operates in a highly competitive market, and the target operates

in a less competitive environment, the deal can generate more value for the combined entity. The coefficient of "GainProdDiff" remains significantly positive at the 5% level. However, the coefficient of "PairSim" is not significant in all regressions of the market reaction analysis. This result suggests that the value creation of the deal may come more from product differentiation rather than the potential reduction of asymmetric information.

In the market reaction analysis of Hoberg and Phillips (2010), they find, for the combined entity, buyers experience higher cumulative abnormal returns when buyers reside in a more competitive industry while targets operate in a less competitive market. My result is consistent with theirs. However, the coefficient of "GainProdDiff" is not significant in their regression when "PairSim" is positively significant. My results differ from theirs.

5.2 Long term performance

In this part, I examine the relationship between post-merger real outcomes and ex-ante product synergy to verify Hypothesis 1 using the following regressions. The dependent variables are profitability scaled by assets and sales, sales growth, and cost changes over one-year and three-year horizons.

$$\text{LT performance}_{i,t} = \beta_0 + \beta_1 \text{ProdSimAcq}_{i,t} + \beta_2 X_{acq,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \tag{5}$$

$$\text{LT performance}_{i,t} = \beta_0 + \beta_1 \text{GainProdDiff}_{i,t} + \beta_2 \text{PairSim}_{i,t} + \beta_3 X_{i,t} + \text{Year F.E.} + \text{Country F.E.} + \epsilon_{i,t} \tag{6}$$

For this analysis, I only focus on the performance of the acquirers and consider the post-merger effective change in performance. Hypothesis 1 suggests that profitability should have a positive relationship with product synergy. In other words, if a target has products similar to those of the acquirer but different from those of the acquirer's close rivals, the post-merger performance should improve due to the full exploitation of product synergy. I examine changes from year t+1 to year t+2 or t+4 (one-year and three-year horizons), resulting in a smaller sample compared to the analysis of the event study. The sample size reduces from 3,017 to 1,220 as I require buyers to have available financial data such as total assets in the following continuous four years after the deal. I construct variables of profitability, sales growth, and cost changes. To reduce the impact of extreme values, I winsorize the profitability, sales growth, and

cost changes variables at the 10% level which means Winsorize the top 5% and bottom 5% of data points. Hoberg and Phillips (2010) winsorize the long-term performance variables at the 1% level. I choose the 10% level because more outliers lie in the data from Datastream. However, Table C shows outliers still exist when checking the maximum or the minimum values of the variables. More work needs to be conducted to deal with this issue.

Table 2 presents the regression results of the profitability variables scaled by assets and sales over one-year and three-year windows. As per the results, I observe that the buyer's profits improve after the transaction if the buyer operates in a competitive market. The coefficient of "ProdSimAcq" is significantly positive at the 1% level with a t-value to be 3.26 for the profitability scaled by total assets over the three-year horizon. "ProdSimAcq" remains significantly positive at the 1% level with a t-value of 4.5 for the profitability scaled by sales over the three-year horizon. However, the coefficient of "GainProdDiff" is no longer significant, but it remains positive. Moreover, the "PairSim" variable has a positively significant coefficient for the three-year horizon at the 1% level with a t-value of 3.04. In the regression of the profitability scaled by assets, Hoberg and Phillips (2010) find the coefficient of "ProdSimAcq" and "PairSim" to be positively significant. In the table of the profitability scales by sales, all three main variables are not significant but positive. Until now, the results of this study are consistent with that of Hoberg and Phillips (2010).

However, when come to the sales growth and cost changes regressions, I find that none of the three main variables has a significant coefficient. Hoberg and Phillips (2010) find a similar result in the cost changes regression but the three main variables are significantly positive for the sales growth regression. The findings of this study suggest that business-similar mergers may not bring asset complementarity but generate other synergies for buyers on the global market.

The results of long-term performance analyses suggest that similar mergers can help firms integrate better after the transaction, particularly when the buyer is in a competitive market. However, the analysis does not detect any significant improvement in sales growth, cost savings, or potential product differentiation. Therefore, the value creation may come from other channels rather than asset complementarity. Based on the positively significant coefficient of the "PairSim" variable, one of the channels could be improved operational efficiency, where the firm can streamline its processes, reduce waste, and optimize its resources to increase efficiency. The firm may also be able to identify and pursue new business opportunities, such as entering new markets, which can lead to revenue growth and increased profitability. It happens when the buyer and seller are from different countries, and owning similar businesses helps the buyer to

enter the seller's market easily.

5.3 Domestic vs Cross-border transactions

To examine if the impact of asset complementarity differs for domestic and cross-border deals, I divide the sample into these two categories domestic and cross-border transactions. To analyze domestic competitors, I establish a criterion that requires countries to have a minimum of 20 transactions during the study period. This selection process yields 12 countries and areas, namely the United States, Canada, Japan, Australia, the United Kingdom, South Korea, India, France, China (Mainland), Taiwan, South Africa, and Sweden. For cross-border deals, I employ a similar criterion, stipulating that countries must have engaged in the purchase and sale of at least 20 foreign firms. Following this criterion, only 4 countries meet the requirement, namely the United States, Canada, the United Kingdom, and Australia.

Table 4 presents the regression results for domestic transactions. The coefficient of "GainProdDiff" is significantly positive across dependent variables and event windows. For instance, in the regression of cumulative abnormal return of buyers, "GainProdDiff" is significantly positive at the 1% level with a t-value of 2.58. Additionally, more value is created for the combined entity when the targets operate in a less competitive market as the coefficient of "ProSimTarg" remains negatively significant at the 5% level. In terms of the improvements in profitability, Table 5 shows that"ProSimAcq" is significantly positive at the 1% level across the two measures over the three-year horizon with t-value of 3.15 and 3.78 respectively. In addition, "PairSim" also remains positively significant at the 5% level across the two measures over the three-year horizon. However, "GainProdDiff" is only positively significant at the 10% level for the profitability scaled by total assets over a one-year horizon. Besides, no improvement in sales growth or cost reduction is observed in Table 6. The results of domestic deals are consistent with the regression results based on the global market.

For cross-border transactions, I add country-level control variables, such as the difference in GDP growth between the buyer's and seller's country. Table 7 shows that all the variables of interest have non-significant coefficients in the event study. However, regarding long-term performance, Table 8 shows that buyers experience an improvement in profitability if the buyers operate in an intensively competitive industry. The coefficient of "ProdSimAcq" is positively significant, especially for the profitability scaled by sales with a t-value of 3.53. However, the "PairSim" variable shows a negative impact on the performance outcomes over the one-year horizon and turns to positive over a longer period. When further digging into the sales growth and

cost reduction analysis, Table 9 shows that "ProdSimAcq" is negatively significant at the 1% level, and "PairSim" is also negatively significant at the 5% level for the cost changes analysis over the one-year horizon. A negative coefficient here means a positive impact on cost reduction, but the impact disappears for a longer period. The "PairSim" variable also has a significant negative coefficient for the short window, which is contrary to my hypothesis that sharing similar businesses between the buyer and seller can generate asset complementarity and help integration. Nevertheless, both coefficients become non-significant for the three-year horizon. These results suggest that business-similar mergers do bring synergies for the buyers but not through the channel of asset complementarity even for cross-border deals.

5.4 Public targets vs Private targets

In their 2006 study, Faccio, McConnell, and Stolin conducted an examination of mergers and acquisitions (M&As) involving both public and private sellers across 17 Western European countries. Their findings revealed a noteworthy trend: the abnormal return for buyers of private targets demonstrated a significant positive effect, whereas that of public targets showed a nonsignificant negative effect. This result is consistent with the findings of the research based on the U.S. firms (see Fuller, Netter, and Stegemoller, 2002; Hansen and Lott, 1996; and Moeller, Schlingemann, and Stulz, 2004). To delve deeper into this phenomenon, Faccio, McConnell, and Stolin(2006) explored various potential factors, including cross-border deals and the introduction of a new blockholder. However, none of these factors could account for the observed listing effect. Consequently, I aim to investigate whether asset complementarity could offer insights into understanding this listing effect.

I collect M&As data worldwide, specifically focusing on deals involving public buyers and private sellers. The information is obtained from Eikon, and I extract the relevant financial data of the buyers from Datastream. In total, there are 17,442 deals analyzed for the announcement return analysis and 10,974 deals for the long-term performance analysis.

Based on the findings presented in Table 10, it is observed that buyers faced a negative abnormal return at the 10% level with the short event window when operating in a competitive environment. However, this effect disappears when the analysis is extended to a wider window. Additionally, the coefficients of the variables "GainProdDiff" and "PairSim" are found to be statistically insignificant.

In terms of the long-term performance of profitability scaled by sales, the regression results of Table 11 indicate that buyers, within a three-year timeframe following the

transactions, exhibit enhanced profitability if they operate in a competitive industry and when the merging pairs exhibit similarity in the product space. "ProdSimAcq" is significantly positive at the 5% level with a t-value of 2.39 and "PairSim" is significantly positive at the 5% level with a t-value of 2.52, both based on the three-year horizon. However, when examining the variables of sales growth and cost reductions, the coefficients associated with the three tested variables are deemed statistically insignificant. Combined with the results from the transactions of public targets, the results suggest that related mergers do bring synergies to the buyers but not through asset complementarity regardless of the public status of targets.

5.5 U.S. M&As

Hoberg and Phillips (2010) focus on analyzing M&A transactions in the U.S. over a span of ten years, from 1997 to 2006. Their objective is to examine the impact of asset complementarity on public M&As. The findings of their study indicate that buyers experience positive abnormal returns during the announcement period and witnessed long-term sales growth through the introduction of new products when the merging pairs exhibit asset complementarity.

To investigate a similar context within the U.S., I specifically consider transactions where both the buyers and sellers originated from the U.S., resulting in a dataset of 1,380 transactions. Table 13 shows that Within a shorter window of surrounding the announcement, the variable "GainProdDiff" demonstrates a significantly positive effect at the 5% level with a t-value of 2.33. However, this effect diminishes when considering an extended window. The coefficient of "ProSimTarg" is significantly negative at the 5% level over two event windows which suggests the combined cumulative abnormal return increases if targets are in a less competitive industry. The sample size reduces to 486 deals for long-term performance analysis and the results are shown in Table 14. What I find in the table is that all three main variables are negatively significant and which is in contrast to the findings of Hoberg and Phillips (2010). Table 5 shows "GainProdDiff" has a negative impact on firms' sales growth over a one-year horizon while "ProdSimAcq" has a positive impact on buyers' cost reduction over a three-year horizon. These results imply that business-similar mergers, in the long term, have a detrimental effect on firms' performance.

There are potential reasons why the findings of this chapter differ from those of Hoberg and Phillips (2010). Firstly, the variance could be attributed to the difference in the time periods studied. Hoberg and Phillips (2010) focus on U.S. transactions from 1997 to 2006, whereas this chapter's study window spans from 2006 to 2018.

Secondly, dissimilarities in the textual data utilized could contribute to the disparities. While Hoberg and Phillips (2010) employ Item 1 of 10-K files, this project relies on TRBD data. Further investigation is necessary to fully understand and explore the discrepancies between the findings of Hoberg and Phillips (2010) and the present study. Thirdly, the outliers in the long-term performance variables of this study may bias the results.

6 Conclusions

Utilizing a unique text-based database, this study investigates the impact of asset complementarity on mergers and acquisitions (M&As) on the global market. The findings suggest that when firms with complementary assets combine under common ownership, value creation can occur. Specifically, mergers that involve similar businesses and occur in competitive industries tend to generate positive market reactions and improve long-term profitability. However, the study also highlights that the value creation in these mergers may not solely stem from asset complementarity but can result from other synergies as well. The research emphasizes the importance of considering business similarity, competitive dynamics, and other factors beyond asset complementarity when analyzing the outcomes of global mergers and acquisitions.

Overall, this study provides insights into the complex relationship between asset complementarity and mergers in the global market. It highlights the need to consider various factors such as business similarity and market competitiveness to understand the impact of these mergers on firm performance. The findings contribute to the existing literature by expanding the analysis beyond the U.S. market and shedding light on the role of asset complementarity in cross-border transactions. By recognizing the multi-dimensional nature of value creation, policymakers and practitioners can make more informed decisions regarding mergers and acquisitions in the global marketplace.

References

[1] Adedeji, A. and Ayoush, M.D., 2017. Are Cross Border Acquisitions More Profitable, or Do They Make Profit More Persistent, than Domestic Acquisitions? UK Evidence. *International Business Research*, 10(6), pp.178-188.

[2] Ahern, K.R., Daminelli, D. and Fracassi, C., 2015. Lost in translation? The effect of cultural values on mergers around the world. *Journal of Financial Economics*, 117(1), pp.165-189.

[3] Anand, J., Capron, L. and Mitchell, W., 2005. Using acquisitions to access multinational diversity: thinking beyond the domestic versus cross-border M&A comparison. *Industrial and Corporate Change*, 14(2), pp.191-224.

[4] Bena, J. and Li, K., 2014. Corporate innovations and mergers and acquisitions. *The Journal of Finance*, 69(5), pp.1923-1960.

[5] Bester, H., 1998. Quality uncertainty mitigates product differentiation. *The RAND Journal of Economics*, pp.828-844.

[6] Cai, Y., Tian, X. and Xia, H., 2016. Location, proximity, and M&A transactions. *Journal of Economics & Management Strategy*, 25(3), pp.688-719.

[7] Deng, P. and Yang, M., 2015. Cross-border mergers and acquisitions by emerging market firms: A comparative investigation. *International Business Review*, 24(1), pp.157-172.

[8] Erel, I., Liao, R.C. and Weisbach, M.S., 2012. Determinants of cross-border mergers and acquisitions. *The Journal of Finance*, 67(3), pp.1045-1082.

[9] Fan, J.P. and Goyal, V.K., 2006. On the patterns and wealth effects of vertical mergers. *The Journal of Business,* 79(2), pp.877-902.

[10] Faccio, M., McConnell, J.J. and Stolin, D., 2006. Returns to acquirers of listed and unlisted targets. *Journal of Financial and Quantitative Analysis*, 41(1), pp.197-220.

[11] Frésard, L., Hege, U. and Phillips, G., 2017. Extending industry specialization through cross-border acquisitions. *The Review of Financial Studies*, 30(5), pp.1539-1582.

[12] Fuller, K., Netter, J. and Stegemoller, M., 2002. What do returns to acquiring firms tell us? Evidence from firms that make many acquisitions. *The Journal of Finance*, 57(4), pp.1763-1793.

[13] Hansen, R.G. and Lott, J.R., 1996. Externalities and corporate objectives in a world with diversified shareholder/consumers. *Journal of Financial and Quantitative Analysis*, 31(1), pp.43-68.

[14] Hart, O. and Moore, J., 1990. Property Rights and the Nature of the Firm. *Journal of Political Economy*, 98(6), pp.1119-1158.

[15] Hoberg, G. and Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10), pp.3773-3811.

[16] Hoberg, G. and Phillips, G., 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), pp.1423-1465.

[17] Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, 39(153), pp.41–57.

[18] Lee, K.H., Mauer, D.C. and Xu, E.Q., 2018. Human capital relatedness and mergers and acquisitions. *Journal of Financial Economics*, 129(1), pp.111-135.

[19] Maksimovic, V. and Phillips, G., 2001. The market for corporate assets: Who engages in mergers and asset sales and are there efficiency gains? *The Journal of Finance*, 56(6), pp.2019-2065.

[20] Moeller, S.B., Schlingemann, F.P. and Stulz, R.M., 2004. Firm size and the gains from acquisitions. *Journal of financial economics*, 73(2), pp.201-228.

[21] Moeller, S.B., Schlingemann, F.P. and Stulz, R.M., 2007. How do diversity of opinion and information asymmetry affect acquirer returns? *The Review of Financial Studies,* 20(6), pp.2047-2078.

[22] Opoku, R.A. and Akorli, P.A., 2009. The preference gap: Ghanaian consumers attitudes toward local and imported products. *African Journal of Business Management*, 3(8), pp.350-357.

[23] Nicholson, R.R. and Salaber, J., 2013. The motives and performance of cross-border acquirers from emerging economies: Comparison between Chinese and Indian firms. *International Business Review*, 22(6), pp.963-980.

[24] Rhodes-Kropf, M. and Robinson, D.T., 2008. The market for mergers and the boundaries of the firm. *The Journal of Finance*, 63(3), pp.1169-1211.

[25] Sheen, A., 2014. The real product market impact of mergers. *The Journal of Finance*, 69(6), pp.2651-2688.

[26] Shimizu, K., Hitt, M.A., Vaidyanath, D. and Pisano, V., 2004. Theoretical foundations of cross-border mergers and acquisitions: A review of current research and recommendations for the future. *Journal of international management*, 10(3), pp.307-353.

[27] Srinivasan, S., 2020. Foreign competition and acquisitions. *Journal of Corporate Finance,* 60, p.101484.

[28] Stiebale, J. and Reize, F., 2011. The impact of FDI through mergers and acquisitions on innovation in target firms. *International Journal of Industrial Organization*, 29(2), pp.155-167.

Table 1: Abnormal return of buyers and combined entity on the global market

	(1) ACAR1	(2) ACAR1	(3) ACAR2	(4) ACAR2	(5) CACR1	(6) CACR1	(7) CACR2	(8) CACR2
ProdSimAcq	0.0247		0.00675		0.0386**		0.0297	
	(1.53)		(0.34)		(2.23)		(1.37)	
ProSimTarg					-0.0396***		-0.0447**	
					(-2.72)		(-2.45)	
GainProdDiff		0.0495***		0.0484**		0.0374**		0.0521**
		(2.73)		(2.18)		(2.27)		(2.53)
PairSim		0.0187		-0.00307		0.0128		0.000450
		(1.42)		(-0.19)		(1.07)		(0.03)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	3017	3017	3017	3017	3017	3017	3017	3017
adj. R^2	0.033	0.034	0.038	0.040	0.048	0.047	0.056	0.057

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ACAR1: the buyer's abnormal return from day -5 to day 5.
ACAR2: the buyer's abnormal return from day -10 to day 10.
CCAR2: the combined abnormal return from day -5 to day 5.
CCAR2: the combined abnormal return from day -10 to day 10.
ProdSimAcq: the competition degree around the buyer.
ProSimTarg: the competition degree around the target.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 2: The changes in profitability of buyers on the global market

	(1) PA1	(2) PA1	(3) PA3	(4) PA3	(5) PS1	(6) PS1	(7) PS3	(8) PS3
ProdSimAcq	0.0289		0.335***		-0.0914		1.508***	
	(0.23)		(3.26)		(-0.24)		(4.50)	
GainProdDiff		0.156		0.130		0.121		0.554
		(1.16)		(1.16)		(0.29)		(1.51)
PairSim		0.0233		0.149*		-0.134		0.799***
		(0.24)		(1.85)		(-0.45)		(3.04)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	1220	1220	1220	1220	1220	1220	1220	1220
adj. R^2	0.022	0.024	0.061	0.055	0.021	0.021	0.089	0.081

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA1: change of profitability scaled by assets from year t+2 to year t+1.
PA3: change of profitability scaled by assets from year t+4 to year t+1.
PS1: change of profitability scaled by assets from year t+2 to year t+1.
PS3: change of profitability scaled by assets from year t+4 to year t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 3: The changes in the costs and sales growth of buyers on the global market

	(1) SG1	(2) SG1	(3) SG3	(4) SG3	(5) CR1	(6) CR1	(7) CR3	(8) CR3
ProdSimAcq	-0.0465		-0.0809		-0.0980		-0.273	
	(-0.63)		(-0.26)		(-1.05)		(-1.42)	
GainProdDiff		-0.0234		0.437		-0.0715		-0.244
		(-0.28)		(1.24)		(-0.67)		(-1.11)
PairSim		-0.0410		0.0471		-0.0288		-0.00805
		(-0.68)		(0.19)		(-0.38)		(-0.05)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	1220	1220	1220	1220	1220	1220	1220	1220
adj. R^2	0.076	0.076	0.088	0.089	0.047	0.046	0.061	0.060

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

SG1: sales growth from year t+2 to t+1.
SG3: sales growth from year t+4 to t+1.
CR1: cost changes from year t+2 to t+1.
CR3: cost changes from year t+4 to t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 4: Abnormal return of buyers and combined entity of domestic deals

	(1) ACAR1	(2) ACAR1	(3) ACAR2	(4) ACAR2	(5) CACR1	(6) CACR1	(7) CACR2	(8) CACR2
ProdSimAcq	0.0179		0.0000802		0.0298		0.0222	
	(1.01)		(0.00)		(1.56)		(0.93)	
ProSimTarg					-0.0350**		-0.0420**	
					(-2.20)		(-2.11)	
GainProdDiff		0.0518***		0.0431*		0.0320*		0.0434*
		(2.58)		(1.76)		(1.75)		(1.91)
PairSim		0.00499		-0.0160		0.00499		-0.00575
		(0.34)		(-0.90)		(0.38)		(-0.35)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	2469	2469	2469	2469	2469	2469	2469	2469
adj. R^2	0.036	0.038	0.040	0.043	0.045	0.045	0.055	0.056

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ACAR1: the buyer's abnormal return from day -5 to day 5.
ACAR2: the buyer's abnormal return from day -10 to day 10.
CCAR2: the combined abnormal return from day -5 to day 5.
CCAR2: the combined abnormal return from day -10 to day 10.
ProdSimAcq: the competition degree around the buyer.
ProSimTarg: the competition degree around the target.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 5: The changes in profitability of buyers of domestic deals

	(1) PA1	(2) PA1	(3) PA3	(4) PA3	(5) PS1	(6) PS1	(7) PS3	(8) PS3
ProdSimAcq	0.114		0.310***		0.224		1.182***	
	(1.00)		(3.15)		(0.62)		(3.78)	
GainProdDiff		0.220*		0.0964		0.433		0.328
		(1.66)		(0.84)		(1.03)		(0.90)
PairSim		0.164*		0.163**		0.308		0.635**
		(1.76)		(2.01)		(1.04)		(2.47)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	934	934	934	934	934	934	934	934
adj. R^2	0.001	0.003	0.036	0.031	0.002	0.002	0.049	0.042

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA1: change of profitability scaled by assets from year t+2 to year t+1.
PA3: change of profitability scaled by assets from year t+4 to year t+1.
PS1: change of profitability scaled by assets from year t+2 to year t+1.
PS3: change of profitability scaled by assets from year t+4 to year t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 6: The changes in the costs and sales growth of buyers of domestic deals

	(1) SG1	(2) SG1	(3) SG3	(4) SG3	(5) CR1	(6) CR1	(7) CR3	(8) CR3
ProdSimAcq	0.0182		-0.0746		0.0168		-0.216	
	(0.21)		(-0.21)		(0.16)		(-0.98)	
GainProdDiff		0.00373		0.505		-0.00625		-0.212
		(0.04)		(1.21)		(-0.05)		(-0.83)
PairSim		-0.0238		0.00280		0.00605		-0.0695
		(-0.34)		(0.01)		(0.07)		(-0.39)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	934	934	934	934	934	934	934	934
adj. R^2	0.059	0.058	0.099	0.100	0.052	0.051	0.065	0.064

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

SG1: sales growth from year t+2 to t+1.
SG3: sales growth from year t+4 to t+1.
CR1: cost changes from year t+2 to t+1.
CR3: cost changes from year t+4 to t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 7: Abnormal return of buyers and combined entity of cross-border deals

	(1) ACAR1	(2) ACAR1	(3) ACAR2	(4) ACAR2	(5) CACR1	(6) CACR1	(7) CACR2	(8) CACR2
ProdSimAcq	0.0622		0.0477		0.0827		0.0490	
	(1.15)		(0.69)		(1.44)		(0.68)	
ProSimTarg					-0.0856		-0.0807	
					(-1.49)		(-1.12)	
GainProdDiff		0.0294		0.0964		0.0543		0.100
		(0.47)		(1.21)		(0.96)		(1.42)
PairSim		0.0621		0.0481		0.0267		-0.0126
		(1.47)		(0.89)		(0.68)		(-0.26)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Country variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	335	335	335	335	334	334	334	334
adj. R^2	0.004	0.004	0.009	0.010	0.019	0.013	0.021	0.026

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ACAR1: the buyer's abnormal return from day -5 to day 5.
ACAR2: the buyer's abnormal return from day -10 to day 10.
CCAR2: the combined abnormal return from day -5 to day 5.
CCAR2: the combined abnormal return from day -10 to day 10.
ProdSimAcq: the competition degree around the buyer.
ProSimTarg: the competition degree around the target.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 8: The changes in the profitability of buyers of cross-border deals

	(1) PA1	(2) PA1	(3) PA3	(4) PA3	(5) PS1	(6) PS1	(7) PS3	(8) PS3
ProdSimAcq	-0.534		0.394*		-1.073		2.450***	
	(-1.21)		(1.81)		(-0.92)		(3.53)	
GainProdDiff		-0.313		0.0992		-0.737		0.647
		(-0.62)		(0.39)		(-0.55)		(0.79)
PairSim		-0.864**		0.109		-1.754*		1.234**
		(-2.47)		(0.62)		(-1.89)		(2.16)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Country variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	158	158	158	158	158	158	158	158
adj. R^2	0.001	0.027	0.062	0.037	0.001	0.013	0.122	0.070

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA1: change of profitability scaled by assets from year t+2 to year t+1.
PA3: change of profitability scaled by assets from year t+4 to year t+1.
PS1: change of profitability scaled by assets from year t+2 to year t+1.
PS3: change of profitability scaled by assets from year t+4 to year t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 9: The changes in the costs and sales growth of buyers of cross-border deals

	(1) SG1	(2) SG1	(3) SG3	(4) SG3	(5) CR1	(6) CR1	(7) CR3	(8) CR3
ProdSimAcq	-0.352		-0.0449		-0.957***		-0.796	
	(-1.53)		(-0.10)		(-3.48)		(-1.53)	
GainProdDiff		-0.0170		0.109		-0.194		-0.780
		(-0.06)		(0.20)		(-0.60)		(-1.29)
PairSim		-0.158		0.423		-0.464**		-0.124
		(-0.85)		(1.14)		(-2.05)		(-0.30)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Country variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	158	158	158	158	158	158	158	158
adj. R^2	0.013	0.030	0.022	0.020	0.065	0.010	0.009	0.002

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

SG1: sales growth from year t+2 to t+1.
SG3: sales growth from year t+4 to t+1.
CR1: cost changes from year t+2 to t+1.
CR3: cost changes from year t+4 to t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 10: Abnormal return of buyers of deals with private targets

	(1) ACAR1	(2) ACAR1	(3) ACAR2	(4) ACAR2
ProdSimAcq	-0.0212*		0.00182	
	(-1.83)		(0.10)	
GainProdDiff		-0.0134		-0.000774
		(-1.17)		(-0.04)
PairSim		-0.0126		-0.00464
		(-1.34)		(-0.32)
Control variables	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓
N	17442	17442	17422	17422
adj. R^2	0.118	0.118	0.130	0.130

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ACAR1: the buyer's abnormal return from day -5 to day 5.
ACAR2: the buyer's abnormal return from day -10 to day 10.
CCAR2: the combined abnormal return from day -5 to day 5.
CCAR2: the combined abnormal return from day -10 to day 10.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 11: The changes in profitability of buyers of deals with private targets

	(1) PA1	(2) PA1	(3) PA3	(4) PA3	(5) PS1	(6) PS1	(7) PS3	(8) PS3
ProdSimAcq	0.0133		0.0150		-0.0499		0.135**	
	(0.65)		(0.72)		(-0.83)		(2.39)	
GainProdDiff		0.0276		-0.0000532		-0.00312		0.0308
		(1.41)		(-0.00)		(-0.05)		(0.57)
PairSim		0.0220		0.0315*		0.0469		0.112**
		(1.36)		(1.91)		(0.99)		(2.52)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	10974	10974	10974	10974	10974	10974	10974	10974
adj. R^2	0.161	0.161	0.106	0.106	0.165	0.165	0.123	0.123

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA1: change of profitability scaled by assets from year t+2 to year t+1.
PA3: change of profitability scaled by assets from year t+4 to year t+1.
PS1: change of profitability scaled by assets from year t+2 to year t+1.
PS3: change of profitability scaled by assets from year t+4 to year t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 12: The changes in the costs and sales growth of buyers of deals with private targets

	(1) SG1	(2) SG1	(3) SG3	(4) SG3	(5) CR1	(6) CR1	(7) CR3	(8) CR3
ProdSimAcq	0.00599		0.0220		0.0434		0.158	
	(0.21)		(0.32)		(1.12)		(1.54)	
GainProdDiff		-0.0270		-0.00319		0.0174		0.0478
		(-0.97)		(-0.05)		(0.47)		(0.49)
PairSim		0.0259		0.0429		0.0320		0.0347
		(1.13)		(0.78)		(1.05)		(0.43)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Country F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	10974	10974	10974	10974	10974	10974	10974	10974
adj. R^2	0.082	0.083	0.077	0.077	0.057	0.056	0.052	0.051

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

SG1: sales growth from year t+2 to t+1.
SG3: sales growth from year t+4 to t+1.
CR1: cost changes from year t+2 to t+1.
CR3: cost changes from year t+4 to t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 13: Abnormal return of buyers and combined entity of the U.S. deals

	(1) ACAR1	(2) ACAR1	(3) ACAR2	(4) ACAR2	(5) CACR1	(6) CACR1	(7) CACR2	(8) CACR2
ProdSimAcq	0.0185		-0.0323		0.0300		0.0186	
	(0.89)		(-1.25)		(1.35)		(0.69)	
ProSimTarg					-0.0371**		-0.0625***	
					(-2.14)		(-2.97)	
GainProdDiff		0.0552**		0.0394		0.0244		0.0380
		(2.33)		(1.34)		(1.13)		(1.45)
PairSim		0.0194		-0.0134		0.0167		0.00471
		(1.16)		(-0.64)		(1.10)		(0.25)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	1380	1380	1380	1380	1380	1380	1380	1380
adj. R^2	0.028	0.031	0.029	0.030	0.068	0.066	0.077	0.072

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ACAR1: the buyer's abnormal return from day -5 to day 5.
ACAR2: the buyer's abnormal return from day -10 to day 10.
CCAR2: the combined abnormal return from day -5 to day 5.
CCAR2: the combined abnormal return from day -10 to day 10.
ProdSimAcq: the competition degree around the buyer.
ProSimTarg: the competition degree around the target.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 14: The changes in profitability of the U.S. deals

	(1) PA1	(2) PA1	(3) PA3	(4) PA3	(5) PS1	(6) PS1	(7) PS3	(8) PS3
ProdSimAcq	-0.360**		-0.290**		-1.301***		-0.889***	
	(-2.15)		(-2.31)		(-2.81)		(-2.64)	
GainProdDiff		-0.165		-0.254*		-0.259		-0.691*
		(-0.87)		(-1.79)		(-0.50)		(-1.82)
PairSim		-0.186		-0.201**		-0.716**		-0.584**
		(-1.51)		(-2.18)		(-2.10)		(-2.35)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	486	486	486	486	486	486	486	486
adj. R^2	0.005	0.002	0.036	0.033	0.008	0.001	0.049	0.044

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA1: change of profitability scaled by assets from year t+2 to year t+1.
PA3: change of profitability scaled by assets from year t+4 to year t+1.
PS1: change of profitability scaled by assets from year t+2 to year t+1.
PS3: change of profitability scaled by assets from year t+4 to year t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Table 15: The changes in the costs and sales growth of the U.S. deals

	(1) SG1	(2) SG1	(3) SG3	(4) SG3	(5) CR1	(6) CR1	(7) CR3	(8) CR3
ProdSimAcq	-0.122		-0.594		-0.127		-0.626*	
	(-1.09)		(-1.56)		(-0.94)		(-1.86)	
GainProdDiff		-0.252**		-0.689		-0.157		-0.531
		(-2.01)		(-1.61)		(-1.03)		(-1.40)
PairSim		-0.126		-0.424		-0.0624		-0.214
		(-1.54)		(-1.51)		(-0.62)		(-0.86)
Control variables	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓
N	486	486	486	486	486	486	486	486
adj. R^2	0.014	0.018	0.026	0.025	0.009	0.007	0.025	0.020

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

SG1: sales growth from year t+2 to t+1.
SG3: sales growth from year t+4 to t+1.
CR1: cost changes from year t+2 to t+1.
CR3: cost changes from year t+4 to t+1.
ProdSimAcq: the competition degree around the buyer.
GainProdDiff: the business distance between the seller and the buyer's ten closest rivals.
PairSim: the business similarity between the buyer and seller.

Chapter 3: Multi-lingual news and stock market returns across countries *

Haim Kedar-Levy †
Xiaojuan Liu ‡
David Stolin §
Fang Xu ¶

Abstract

Despite the central importance played by news in driving security prices, there has been no large-scale work examining the price impact of news across countries and languages. A critical barrier is different languages and the need to analyse the news content. To address this gap, we study how news in several languages explain returns on several dozen countries' stock market indices. Specifically, we use a comprehensive dataset of more than 270 million worldwide online news articles together with a powerful and novel multi-lingual text processing technology to assess the value of news relevant in each country. We develop a "news intensity" variable and study its incremental contribution to explaining the volatility of the country's stock market index. Our results show that news coverage in different languages impacts stock markets around the globe. Generally, a greater volume of relevant news tends to predict a dampening of volatility, consistent with news coverage serving to resolve rather than generate uncertainty. Results across languages are harder to interpret, with languages other than English or the local language frequently having a significant impact. Overall, or study points to both the promise and challenges of working with textual big data in the investment setting.

Keywords:news, volatility, stock markets, natural language processing

*We are grateful to Jasper Ginn, Crina Pungulescu, Francisco Webber and Maxim Zagonov for many insightful conversations. Special thanks to Jasper Ginn for extensive assistance with the data. All errors are ours.

†Ben Gurion University, Beer Sheba, Israel: hlevy@som.bgu.ac.il
‡Toulouse School of Management, Toulouse, France: xiaojuan.liu@tsm-education.fr
§Toulouse Business School, Toulouse, France: d.stolin@tbs-education.fr
¶Brunel University, London, UK: fang.xu@brunel.ac.uk

1 Introduction

Understanding how news coverage impacts asset prices is of central importance for investors, investment managers, regulators, and researchers. The reason is that news coverage directs investor attention, which in turn has been shown to affect price discovery and volume (e.g., Fang and Peress 2009, Engelberg and Gao 2011, Baloria and Heese 2018). However, with the notable exception of Byström (2016)[1] , there has been no comparative cross-country (let alone cross-language) work on the impact of news coverage.

In order address this gap, we employ a novel dataset combined with a novel textual analysis technology. Specifically, we use a comprehensive dataset of millions of online news stories from across the world in five languages (Arabic, French, English, German and Russian). Our goal is to examine the impact of daily news volume on the volatility of stock market index returns for a broad range of countries. In order to quantify the volume of news relevant to each country we use the semantic fingerprinting approach validated in the investment context by Ibriyamova et al. (2017, 2019). Our key results report on the significance of relevant news volume (for news in different languages) in helping predict country stock market index volatilities.

The efficient markets hypothesis, which is of central importance to financial economics, states that asset prices should reflect public information fully and rapidly. Indeed, the notion that information contained in public news does impact asset prices has found empirical support in numerous studies (Bali, Bodnaruk and Scherbina, 2017; Boudoukh et al., 2018; Heston and Sinha, 2017; Boukus and Rosenberg, 2006, Calomiris and Mamaysky, 2018, Engelberg and Parsons, 2011; Glasserman and Mamaysky, 2018; Jiang, Li and Wang, 2017; Larsen and Thorsrud 2015, Box, 2018). In contrast to the efficient market hypothesis, however, most of these studies find that the economic content of news is not incorporated into prices fully and instantaneously, as both under- and over-reaction is present. Even more damningly for market efficiency, a number of papers document that news coverage impacts asset prices even if it is devoid of new and informative content (Da et al., 2011; Fang and Peress, 2009; Fedyk and

[1]To our knowledge, Byström (2016) is the only other paper that examines the impact of news coverage across languages and countries. Working at the monthly frequency, he shows stock-market related news coverage in English and Chinese to predict volatility of Global, US, UK and Chinese market indices. There are a number of important differences between our approach and Byström's. First, our data frequency is daily. Second, we examine five languages instead of two, and 35 countries instead of three. Third, rather than screening for news that mention the stock market, we screen for country-relevant news using a sophisticated NLP algorithm.

Hodson, 2017; Huberman and Regev, 2001; Solomon, 2012; Tetlock, 2010, Loughran and McDonald, 2016). The notion that news coverage (as distinct from news content) influences asset prices is predicated on it focusing investor attention (Barber and Odean, 2008). In a striking recent paper, Cohen et al. (2018) even report that news content (represented by changes in the text of 10-K filings) can have no immediate effect on company returns in the absence of news coverage.

While studying the link between news and asset prices is a fundamental issue in finance, the scope of almost all of the above studies is limited in three important regards. First, the examined news stories have information directly referencing the asset studied (generally, individual corporations and their share prices). However, asset prices are also liable to be impacted by relevant news that does not reference the asset directly. Second, the assets studied in the literature, with very few exceptions, are US-based. Yet, this raises significant questions about the generalizability of findings in a world where less than half of the global stock market cap is US-based (Dimson, Marsh and Staunton, 2018). Third, the news stories examined in the literature are overwhelmingly in English. This is at odds with the reality of an increasingly globally interrelated economy, where much news originates from and/or is amplified by non-English news sources.

Our proposed research moves beyond these limitations by conducting a comprehensive study of the impact of indirect news coverage across multiple languages and countries. Our approach is to take the totality of online news published since 2016 and scraped by Webhose.io, together with textual descriptions of individual countries, and to estimate the quantity of (possibly indirectly) related news coverage for each country, each day, and in each language. We then use this quantity (which we refer to as the "news intensity" for a given asset) to help predict the volatility of the corresponding stock market return. With over 270 million news articles analysed, the size of our news dataset exceeds that of ones used in previous studies by at least an order of magnitude.

Our ability to expand the scope of our investigation into news and asset prices to news coverage that is indirect, global and not limited to the English language is due to our reliance on a novel technology called semantic fingerprinting, offered by Cortical.io, which enables efficient capture and comparison of the semantic meaning of texts both within and across languages (Webber 2015; Ibriyamova et al. 2017, 2019). Together, these innovations mean that for a news story in any of Cortical.io-supported languages, we can calculate its contribution to the 'news intensity' of any asset (a measure of the

volume of related news coverage) via the overlap between the semantic fingerprints of the news story and that of the asset in question. In the case we address in this chapter, that of country stock market indices, we use semantic fingerprints of individual countries.

Our key results are threefold. First, news intensities consistently prove to be significant for forecasting stock index volatility – much more frequently than would be expected solely due to chance – both in in-sample and in out-of-sample tests. Second, when news intensities do have a significant impact, this impact is generally negative, suggesting that greater online news coverage is associated with a calming effect on investors and markets. Third, the role of the language of the news is ambiguous, with international stock markets often being impacted by news in a language that is neither English nor the local language. Overall, our results indicate that textual big data, the so-called "big text", holds substantial promise for insight into the behavior of asset prices, although the very richness of the data is also liable to present challenges of interpretation.

The rest of the chapter is organized as follows. The next section presents our methodology while Section 3 overviews our data. Section 4 presents out results. Section 5 concludes.

2 Methodology

Our working hypothesis is straightforward: we expect that related news coverage has an impact on volatility. In particular, it is natural to expect that the more news coverage there is, the greater the volatility (e.g. Byström 2016). However, one can also argue that an opposite effect may be at play: increased news coverage may have a calming effect on investors, e.g. if it is seen as resolving uncertainty about the future or if the prevalent economic narrative (Shiller 2017) is seen to be of a reassuring nature. Therefore, the direction of any effect remains an empirical issue.

In order to operationalize our tests of the impact of multi-lingual news on stock returns across countries, we require the following elements:

1. An efficient approach to processing a large body of news text. We do so using the semantic fingerprinting methodology.

2. Once the news text has been processed, we need a measure capturing the volume of relevant news. We call this measure news intensity.

3. Once news intensities have been obtained, we need an econometric approach to study their impact on returns.

Semantic fingerprinting, news intensities and our econometric approach are described below.

Semantic fingerprinting

Semantic fingerprinting as implemented by Cortical.io (Webber 2015) represents text as a sparse binary 128x128 matrix of 16,384 topics. The semantic fingerprint of a word is represented by topics with which the word is associated in a training corpus. The semantic fingerprint of a text is the aggregation of semantic fingerprints of key words in that text. Ibriyamova et al. (2017, 2019) demonstrate semantic fingerprinting to be an effective document comparison tool in the investment context. In particular, they show that similarity of company descriptions, measured as the cosine similarity of their semantic fingerprints, predicts the similarity of their stock returns. In this project, we take semantic fingerprints of the countries in our sample, and use them as filters for the semantic fingerprints of news stories, in order to focus on news most relevant to each country. In addition to picking up explicit geographic references in news stories, this approach attributes to a country news stories that are likely to be of relevance even if the country itself is not mentioned in them; for example, the semantic fingerprints of oil-producing countries will have greater overlap with the semantic fingerprints of oil-related news. In this way, we operationalize the volume of relevant news as the news intensity for any country/language/day combination.

News intensities

More formally, let $N(t)$ be the total number of news stories on day t. Corresponding to each news story is a semantic fingerprint comprising a subset of K possible positions, each of them representing a group of related terms. We define I_i as a K-vector containing the semantic fingerprint of the i-th country. Further, let J_i be a K-vector containing the semantic fingerprint of the nth news story on day t. Then the news intensity score for country i on day t is defined as the average of the number of news stories whose fingerprints include a given position, taken across all positions comprising country i's fingerprint, and scaled by the average number of stories across

all possible positions:

$$s_{i,t} = \frac{\left(\sum_{n=1}^{N(t)} I_i \cdot J_{n,t}\right) / \|I_i\|}{\sum_{n=1}^{N(t)} \|J_{n,t}\| / K}$$

As a simple illustration, suppose that we use a semantic space comprised of only $K = 2 \times 2 = 4$ positions (instead of $K = 128 \times 128 = 16384$ we actually employ in our analyses [2]). Suppose further that Asset A's fingerprint is $(1, 1, 0, 0)$, and Asset $B's$ is $(0, 1, 1, 1)$, and that on Day t there are $N(t) = 5$ news stories whose fingerprints are $(1, 0, 0, 0), (1, 1, 0, 0), (1, 1, 1, 0)$ and $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$. Then the aggregate semantic fingerprint across the five news stories is $(4, 3, 2, 1)$ and the total number of semantic positions across the four news stories is $\sum_{n=1}^{N(t)} \|J_{n,t}\| = 4+3+2+1 = 10$. Of them, $\sum_{n=1}^{5} I_A \cdot J_{n,t} = 7$ positions overlap with A's fingerprint, and $\sum_{n=1}^{5} I_B \cdot J_{n,t} = 6$ positions overlap with B's fingerprint. So the number of relevant news fingerprint positions per Asset A fingerprint position is $\left(\sum_{n=1}^{N(t)} I_A \cdot J_{n,t}\right) / \|I_i\| = \frac{7}{2} = 3.5$; the analogous quantity for Asset B is $\left(\sum_{n=1}^{N(t)} I_B \cdot J_{n,t}\right) / \|I_i\| = \frac{6}{3} = 2$. Dividing these by the scaling factor of $\sum_{n=1}^{N(t)} \|J_{n,t}\| / K = 10/4 = 2.5$ produces news intensities of $s_{A,1} = \frac{3.5}{2.5} = 1.4$ and $s_{b,2} = \frac{2}{2.5} = 0.8$. Their interpretation is that the amount of news per fingerprint position for Asset A is 40% higher and for Asset B, 20% lower than the average amount of news for a randomly chosen fingerprint position.

Note that the above news intensity score is constructed separately for news in each of the six languages we use in our study. We distinguish scores for different languages via superscripts I, i.e. news intensity scores in different languages are denoted $s^l_{i,t}$, where $l = 1, \ldots, 6$.

Econometric approach

We use two approaches to investigate the contribution of the news intensity to the volatility of country index returns, following similar methods discussed in Day and Lewis (1992). First, we study the contribution of the news intensity to the conditional volatility of the stock market index returns via an in-sample analysis. We add the news intensity as an exogenous variable to exponential GARCH model (Nelson, 1991) for

[2] The keywords related to each position can be obtained via http://languages.cortical.io/Expression.htm#!/expressions/getSimilarTermsForSinglePosition_get_6. Note that, consistently with a convention used by many programming languages, Cortical numbers the positions starting from 0 and not from 1 .

the index returns. The EGARCH model takes into account the asymmetric effects of positive and negative shocks on volatilities, compared to the standard GARCH model (Bollerslev 1986; Taylor 1986). Our data confirm that the asymmetric effects are significant for most countries. The EGARCH model also doesn't impose non-negativity constraints on the model parameters, compared to the GJR model (Glosten, Jagannathan, and Runkle 1993). Second, we study the out-of-sample predictive content of news intensity for ex post volatility of index returns, along with rival forecasts based on historical volatilities and the EGARCH model.

For the EGARCH model, the mean equation for the excess return can be written as

$$R_{Mt} - R_{Ft} = \lambda_0 + \varepsilon_t \tag{1}$$

where R_{Mt} is the return of the country stock index, R_{Ft} is the risk-free rate approximated by the short-term interest rate for the corresponding country, λ_0 is a constant coefficient, and ε_t is a random error term that is normally distributed with mean zero and variance σ_t^2.

$$\ln\left(\sigma_t^2\right) = \omega + \beta \ln\left(\sigma_{t-1}^2\right) + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha \left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| \tag{2}$$

This equation models the conditional variance in natural logarithm ($\ln(\sigma_t^2)$) as a function of the previous conditional variance in natural logarithm ($\ln\left(\sigma_{t-1}^2\right)$), the previous standardized shock ($\varepsilon_{t-1}/\sigma_{t-1}$), and its absolute value ($|\varepsilon_{t-1}/\sigma_{t-1}|$). The inclusion of both the standardized shock and its absolute value captures the asymmetrical contribution of the positive and negative shocks of equal magnitude to volatility. If the relationship between volatility and returns is negative, then γ will be negative. Also, as the natural logarithm of variance instead of the variance itself is modelled in equation (2), the parameters' values don't have the non-negative constraints. Even if the parameters are negative, the conditional variance (σ_t^2) will be positive.

Our results from the above EGARCH$(1,1)$ are robust. We have considered a range of alternative specifications, which don't provide qualitatively different results. For example, we have considered generalized error distribution (instead of the normal distribution) for the errors, as originally proposed by Nelson (1991). Results from this are similar. We have also considered EGARH-in-mean specification, with the standard error σ_t as an additional regressor in the mean equation (1). However, this additional regressor is rarely significant for our considered data. The effect from the conditional

variance (standard deviation) to the mean returns cannot be confirmed.

For the in-sample (IS) analysis, we examine the information content of the news volume by adding the news intensity (s_t) as an exogenous explanatory variable to the conditional variance equation, i.e.

$$\ln\left(\sigma_t^2\right) = \omega + \beta \ln\left(\sigma_{t-1}^2\right) + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha \left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| + \delta s_{t-1} \tag{3}$$

The coefficient δ can be interpreted as a measure of the incremental information which news intensity in the previous period contributes to changes in the returns variance in this period. Thus, the hypothesis that the news intensity impounds information in addition to what can be obtained from the historic series of returns can be tested by examining the statistical significance of δ. To test the null hypothesis $\delta = 0$, we consider both the t-test and the likelihood ratio (LR) test. The tstatistic is based on standard errors using the robust inference procedures proposed by Bollerslev and Wooldridge (1992). For the LR test statistic, the unrestricted model is equation (3) and the restricted model is equation (2).

For the out-of-sample (OOS) analysis, we investigate the relative predictive power of the news intensity to the one-step ahead return volatility. We use the square of the daily return on the index as the proxy for ex post volatility. First, we consider a series of regressions of the following form:

$$\sigma_{t+1}^2 = b_0 + b_1 \sigma_{Ft}^2 + \epsilon_{t+1} \tag{4}$$

where σ_{t+1}^2 is the ex-post volatility during period $t+1$, σ_{Ft}^2 is a forecast of volatility at the end of period t, and ϵ_{t+1} is the forecast error. We use news intensity, historical volatility (volatility in the previous period) and the EGARCH model to form the forecast (σ_{Ft}^2). For the news intensity, the forecast is based on a regression of return volatility on a constant and one-lagged news intensity. Comparing the performance of these three predictive regressions will shed light on the individual predictive power of each model.

Second, we look at the relative predictive power of alternative forecasts via the encompassing tests. We consider all three forecasts simultaneously in the same predictive regression, i.e.

$$\sigma_{t+1}^2 = b_0 + b_1 \sigma_{St}^2 + b_2 \sigma_{Et}^2 + b_3 \sigma_{Ht}^2 + \epsilon_{t+1} \tag{5}$$

where $\sigma^2_{St}, \sigma^2_{Et}$ and σ^2_{Ht} are the forecasts of volatility based on the news intensity, the EGARCH model, and historical volatility accordingly. If the information contained in the news intensity is relevant and different from the information in the EGARCH and historical volatility, the estimate for b_1 shall be non-zero.

3 Data

Our dataset combines two elements: stock market index returns for multiple countries and online news articles in multiple languages.

The countries in our sample are those whose stock market index returns are available from WRDS: Australia, Austria, Belgium, Brazil, Chile, China, Colombia, Denmark, Finland, France, Germany, Hong-Kong, India, Indonesia, Ireland, Italy, Japan, Malaysia, Mexico, the Netherlands, New-Zealand, Norway, Philippines, Poland, Singapore, South-Africa, South-Korea, Spain, Sweden, Switzerland, Taiwan, Thailand, Turkey, the UK, and the US.

The news data come from Webhose.io, and consist of news items scraped from online news sites starting in February 2016 and ending in October 2018. Every news site is scraped at least once a day, with the most popular sites scraped multiple times a day (however, multiple occurrence of the same story are tagged as such, and we exclude such repeats from our dataset). Each news data point comes in the form of an XML file and includes the text of the news item, a time stamp for when it was scraped, item type (we only use items of 'Mainstream' type, i.e. those from recognized news sites as opposed to blogs or other websites) and language and country codes. Appendix A shows a sample news item, and Appendix B shows its semantic fingerprint. A free Webhose.io API is available on $https://webhose.io/products/news-feeds/$.

Table 1 shows the distribution of news stories by language. As expected, English is the dominant news language, with 183,982 different news stories on average per day, with a minimum of 61,574 and a maximum of 252,315. However, all of the other languages also have thousands of stories per day, from a low of 15,635 per day in French to a high of 28,412 per day in Russian. This brings the average daily number of stories across all of our languages in aggregate to 274,056, corresponding to a total number of news stories in our 988-day sample period of more than 270 million, dwarfing the size of news repositories examined by other studies.[3]

[3]For example, one of the largest news samples hitherto examined by researchers, that of Tao,

Table 1. The distribution of news stories across days and languages

	Arabic	English	French	German	Russian	Aggregate
Mean	18,544	183,982	15,635	27,484	28,412	274,056
Standard deviation	3,971	35,515	3,145	7,431	8,211	50,213
Minimum	4,719	61,475	3,881	8,593	7,374	86,043
25th percentile	16,185	158,533	13,266	22,095	21,111	235,825
50th percentile	18,519	191,575	15,639	27,150	29,510	285,028
75th percentile	21,143	211,254	17,993	31,594	35,401	314,642
Maximum	29,766	252,315	30,477	76,812	49,710	372,870

Table 2. Correlations between news intensities

	Arabic	English	French	German	Russian	Aggregate
Arabic	1.000	0.272	0.195	0.228	0.143	0.287
English	0.272	1.000	0.666	0.509	0.572	0.942
French	0.195	0.666	1.000	0.476	0.569	0.731
German	0.228	0.509	0.476	1.000	0.514	0.644
Russian	0.143	0.572	0.569	0.514	1.000	0.669
Aggregate	0.287	0.942	0.731	0.644	0.669	1.000

Given that the existing literature overwhelmingly focuses on news in English, it is especially important to compare the similarity of news coverage across languages. To this end, Table 2 reports correlations for news intensities obtained for pairs of languages (these are first calculated for news intensities pertaining to each country in our sample, and then averaged across countries). While English-language news predictably has a high correlation with news in aggregate (of which English news are the dominant constituent), it has markedly lower correlations with news in French (0.666), Russian (0.572), German (0.509) and especially Arabic (0.272). More generally, low correlations for pairs of languages suggest that monolingual studies of the impact of news on asset prices are unlikely to be generalizable on a global basis. These low correlations also foreshadow the diversity of results across languages in our subsequent regressions.

Brooks and Bell (2019), contains approximately 11 million news stories, spread over the 1979-2016 period.

4 Results

Consistently with the richness of our multi-lingual, cross-country dataset, our estimation results comprise a large number of regressions: separately for each language, country and empirical specification. In order make these results easier to understand, we first present in-sample regressions for a single country, the UK.

Table 3. In-sample estimates for the UK stock market

Mean equation	Variance equation				
λ	ω	β	γ	α	δ
EGARCH without news intensity					
0.026	**-0.163**	**0.952**	**-0.148**	0.154	
EGARCH with news intensity					
Aggregate					
0.027	0.713	**0.957**	**-0.147**	0.153	-0.184
Arabic					
0.023	**-1.503**	**0.954**	**-0.163**	0.135	0.243
English					
0.029	1.596	**0.959**	**-0.150**	0.151	-0.371
French					
0.026	0.388	**0.956**	**-0.151**	0.150	-0.115
German					
0.038	**-8.875**	**0.866**	**-0.193**	0.079	**1.936**
Russian					
0.027	-0.772	**0.950**	**-0.145**	0.152	0.133

Note: This table presents the estimates of the Equation (2) and (3). Coefficients significant at 10% level via the t-test is in boldface. The relevant t-statistics are based on the robust inference procedures by Bollerslev and Wooldrige (1988). Estimates for δ are also in blue if the corresponding LR test rejects the null hypothesis that $\delta = 0$ at 10% level.

Table 3 shows the results of estimating Equations (2) and (3). γ captures the asymmetric reactions of conditional volatility to standardized shocks. Our table shows γ is significantly negative at the 10% level across the languages. The negative relation between the conditional volatility and the standardized shock, which is also found

in Nelson (1988) and Day and Lewis (1992), implies that the magnitude of volatility change triggered by the shock is higher when the shock is negative. When the news intensity is included as the external regressor in the variance equation, German-language news contributes to the conditional volatility significantly as both the t-test and the LR test confirm. In addition, both Arabic and English news also show significant impacts according to the LR test. Note results from the t-test and the LR test can differ. For the null hypothesis of our interest ($\delta = 0$), the LR test compares the maximised likelihood value of the unrestricted model (Equation (3)) with that of the restricted model (Equation (4)). In contrast, the t-test is in fact a Wald test, which is calculated using the unrestricted mode (Equation (3)) only.

Table 4 table presents estimated coefficients of news intensity in the variance equation of conditional volatility of stock returns for all considered countries across all languages. Many of these coefficients are significant either via the t-test or the LR test. There is strong evidence suggesting that news intensities have a significant impact on the conditional volatility. In terms of the sign of the impact, however, the evidence in Table 4 is mixed. The impacts are mostly positive if German-language news is considered, and show mixed results when other languages are used.

The impact of news intensity on the conditional volatility is also time varying. Table 5 shows estimated coefficients of news intensity in the variance equation for two sub-periods: first 300 observations and the remaining more than 200 observations. The division of the period is chosen such that the second period corresponds to the forecasting period in the OOS analysis. As such the IS results from the second period and the OOS analysis provide evidence on the predictive power of the news intensity for the same sample period. Again, the significant impact of the news intensity is confirmed. Many of the coefficients are in boldface (i.e. significantly different from zero via either the t-test or the LR test). To visualize the direction of the impact, we use two colours for these significant coefficients: green for the positive coefficients and red for the negative coefficients. Consider German-language news as an example. For the first sub-period, news intensities have significant impacts on the conditional volatility for 21 out of the 35 countries, and for all these 21 countries news intensities show a positive impact. For the second sub-period, however, news intensities have significant impacts only for 11 out of 35 countries, and for 6 of the 10 countries news intensities show a positive impact. So the overall message from the IS analysis is that news intensities affect the conditional volatility, its effect is time varying, and the direction of the effect can be positive or negative.

Next, we look at the out-of-sample forecasting performance of news intensity in comparison to forecasts based on historical volatility and the EGARCH model. Using

Table 4. In-sample regressions across languages and countries

	Aggregate	Arabic	English	French	German	Russian
Australia	2.297	**3.754**	-2.004	-4.127	**8.405**	2.300
Austria	0.586	0.207	0.632	0.967	**1.550**	**0.680**
Belgium	-1.147	0.390	-0.977	0.492	0.874	0.166
Brazil	-1.914	**1.573**	**-1.734**	**-1.850**	0.598	**-1.559**
Chile	1.352	-0.118	0.824	0.441	**-1.830**	**0.978**
China	0.170	-0.111	0.159	0.181	0.346	0.124
Colombia	-0.969	0.072	-0.181	**-1.367**	-0.829	**-1.536**
Denmark	-0.690	0.577	-0.692	-0.163	**1.486**	-0.103
Finland	1.141	0.302	0.877	1.504	**2.858**	0.675
France	-0.222	0.045	-0.496	0.737	1.190	0.451
Germany	-0.127	0.014	-0.176	0.487	1.261	0.285
HongKong	**0.406**	0.338	**0.228**	-0.130	0.462	-0.723
India	**-0.799**	**0.527**	**-0.675**	-0.341	0.881	**-0.963**
Indonesia	-0.047	-0.183	0.147	-0.586	1.115	-0.252
Ireland	1.076	0.133	1.099	0.264	2.467	0.446
Italy	0.534	0.380	0.662	0.617	**2.840**	0.446
Japan	-0.502	0.256	-0.849	-0.383	0.518	0.071
Malaysia	0.092	-0.064	0.272	-0.330	-0.503	-0.350
Mexico	-1.550	**0.971**	-1.287	-0.835	0.433	-1.375
Netherlands	0.112	0.206	0.096	0.286	1.360	0.333
New Zealand	0.504	0.680	-0.279	-0.888	**3.049**	-0.043
Norway	-0.383	-0.051	**-0.476**	0.090	0.037	0.008
Philippine	**-1.963**	-0.163	**-1.608**	0.026	-0.184	**-0.694**
Poland	-0.203	0.551	0.178	0.618	0.648	0.369
Singapore	0.858	1.538	0.371	-0.064	1.263	-0.101
South Africa	0.431	0.294	0.333	0.145	0.739	0.196
South Korea	0.222	**-0.291**	0.171	0.104	0.059	**0.306**
Spain	-0.345	-0.880	-0.354	0.350	**3.051**	0.358
Sweden	0.679	0.076	0.582	1.301	**2.262**	0.539
Switzerland	0.098	-0.179	0.007	0.789	**1.506**	0.370
Taiwan	-0.584	-0.180	-0.470	-0.347	0.266	-0.150
Thailand	-0.589	**1.891**	-0.930	**0.330**	**1.243**	**-1.962**
Turkey	4.781	2.031	4.209	2.807	**3.944**	2.116
UK	-0.184	0.243	-0.371	-0.115	**1.936**	0.133
US	0.520	0.069	0.362	-0.089	0.754	0.135

Note: This table reports the estimate of δ in equation (3) for all countries and languages. Coefficients that are significantly different from zero at 10% according to the t-test are in boldface and according to the LR test are in blue. The relevant t-statistics are based on robust inference procedures by Bollerslev and Wooldrige (1988).

Table 5. In-sample regressions for two sub-periods

	First Period						Second Period					
	Aggregate	Arabic	English	French	German	Russian	Aggregate	Arabic	English	French	German	Russian
Australia	3.499	1.942	-1.397	-4.261	**7.141**	3.144	-0.326	**1.505**	-0.664	**-3.426**	**12.236**	1.873
Austria	**0.847**	**-0.261**	**6.770**	**0.818**	**1.942**	**0.668**	**-9.547**	**-2.208**	**-12.107**	**-4.640**	-0.628	-1.553
Belgium	-0.412	**-0.492**	-0.861	**0.421**	**0.983**	**0.647**	**-5.231**	-0.170	**-5.695**	-1.507	**-2.593**	-1.316
Brazil	-1.818	0.939	-1.482	-1.690	-0.119	-1.384	**3.512**	**1.994**	**2.502**	-0.094	1.047	0.295
Chile	0.595	-0.030	0.683	0.349	-0.782	0.366	2.553	0.229	-0.263	**-3.201**	-2.498	-1.771
China	**0.024**	**-0.611**	**-0.059**	**-3.142**	**0.049**	**0.019**	**1.693**	0.586	**1.472**	-0.365	0.626	-0.011
Colombia	**-4.849**	-0.653	**-3.335**	**-2.434**	-0.639	**-2.720**	0.655	0.183	1.628	-0.121	-0.330	**-1.245**
Denmark	-0.335	0.131	-0.624	0.409	**2.027**	0.183	-0.403	0.892	0.011	**-1.794**	-0.217	2.023
Finland	**2.219**	-0.153	**1.484**	**2.032**	**3.283**	**1.050**	-1.808	0.020	-0.008	**-4.509**	-0.674	0.526
France	0.402	**-0.624**	-0.259	**1.243**	**1.620**	**0.752**	**-3.632**	0.237	**-2.420**	-0.666	**-3.193**	**-0.987**
Germany	0.227	**-0.209**	0.027	**1.054**	**1.423**	**0.414**	-0.458	0.009	**1.850**	-0.254	-0.809	-0.411
HongKong	0.353	**-0.215**	-0.447	0.233	0.726	0.040	2.284	**1.727**	**3.473**	-1.155	-0.435	**-2.079**
India	**-0.864**	0.361	**-0.235**	-0.087	**0.141**	**-0.639**	-0.930	0.245	-0.882	-0.111	-1.013	-1.129
Indonesia	0.333	-0.491	0.201	-0.144	**1.578**	0.088	0.227	**-1.791**	**2.838**	-1.201	**2.977**	-1.779
Ireland	**2.194**	**-4.023**	**14.425**	**1.363**	**3.235**	**1.342**	**-0.528**	**-0.338**	**-0.702**	**-0.706**	-1.557	**-0.850**
Italy	**1.760**	-0.436	0.997	**1.414**	**3.173**	**0.945**	**-3.153**	0.913	**-2.912**	0.166	-0.647	0.041
Japan	-0.348	-0.099	**-0.946**	0.049	0.587	0.441	0.788	0.983	1.413	-0.540	-0.037	-0.943
Malaysia	0.183	**-0.158**	0.180	0.165	0.245	0.128	**4.276**	**1.402**	**4.809**	-0.087	0.769	-0.352
Mexico	**-2.587**	0.974	**-2.097**	-1.355	-0.664	**-1.977**	-1.096	0.445	-0.770	-0.990	0.250	-1.781
Netherlands	**0.798**	**-0.338**	0.528	**0.560**	**1.639**	**0.288**	-1.320	0.442	0.318	**-1.184**	**-1.701**	-0.170
New Zealand	0.674	0.373	-0.006	0.110	**1.050**	0.004	0.969	-0.097	3.412	0.614	4.011	0.482
Norway	0.037	0.025	0.037	0.054	0.049	0.039	**-5.018**	-1.042	**-3.979**	-1.636	**-3.583**	-0.798
Philippine	**-1.029**	**-0.703**	**-1.000**	0.169	-0.130	**-0.903**	**-7.178**	**-1.951**	**-6.991**	-0.412	**-4.176**	**-0.960**
Poland	1.375	-0.150	0.861	**3.093**	**0.897**	**1.057**	-5.923	**0.290**	**-16.267**	-0.204	-0.803	-0.295
Singapore	-0.910	-2.574	-1.644	-1.349	0.500	-1.348	**2.688**	**1.324**	**3.704**	**-0.575**	0.233	**-1.508**
South Africa	0.033	0.034	0.032	0.041	0.055	0.035	**2.827**	**0.597**	**4.378**	**-3.035**	**6.446**	**6.975**
South Korea	**-3.717**	**-4.478**	**-3.158**	**-3.890**	-1.049	0.813	**1.871**	0.535	**1.675**	-0.058	**1.954**	0.198
Spain	-2.291	**-4.054**	2.507	**3.708**	0.242	-0.820	-1.191	-0.740	**-2.115**	-0.089	0.259	-0.494
Sweden	**1.067**	**-0.372**	0.286	**1.743**	**3.638**	**0.699**	0.267	**0.844**	**1.385**	-0.541	**1.533**	0.137
Switzerland	0.297	**-0.577**	-0.052	**0.916**	**3.193**	**0.523**	**-1.944**	-0.139	-1.577	**-1.451**	-1.308	-0.341
Taiwan	-0.533	-0.153	**-0.983**	**0.475**	**0.363**	-0.053	-0.425	0.648	0.399	**-1.469**	-0.151	-0.041
Thailand	**0.900**	0.356	0.010	**0.651**	**1.005**	**-1.135**	1.034	**2.942**	1.407	-0.056	-0.501	**-3.205**
Turkey	**5.109**	**3.827**	**5.083**	**4.337**	**5.427**	**2.726**	0.325	**-0.014**	**-0.021**	**0.050**	-0.229	**-0.035**
UK	0.281	-0.326	-0.479	**0.613**	**2.979**	**0.450**	0.508	0.297	**1.016**	-0.316	-0.184	0.282
US	-1.698	-0.486	**-1.470**	**-1.508**	0.444	-0.143	**1.445**	**0.810**	**1.831**	-0.348	**1.282**	0.020

Note: Coefficients that are significantly different from zero at 10% according to either the t-test or the LR test are in boldface. In addition, for these significant coefficients, they are in green if positive and in red if negative.

Table 6. Out-of-sample predictive power of alternative forecasts for the UK

$\sigma^2_{t+1} = b_0 + b_1\sigma^2_{Ft} + \epsilon_{t+1}$

	b_0	b_1	R^2
Historical Volatility	**0.198**	**0.400**	0.159
EGARCH	0.010	**1.094**	0.039
S_Aggregate	**0.375**	**-0.136**	0.004
S_Arabic	**0.636**	**-0.962**	0.021
S_English	**0.385**	-0.160	0.004
S_French	**0.376**	-0.135	0.003
S_German	**0.357**	-0.093	0.001
S_Russian	**0.400**	**-0.192**	0.008

Note: Significant coefficients according to the t-test at 10% are in boldface. The relevant t-statistics use the White (1980) corrected standard errors.

the UK as an example, Ta-ble 6 reports estimation results from the individual forecast (Equation (4)) in the OOS analysis. As has been seen in the literature, the historical volatility has the strongest performance, in our case with an R-squared of 0.159. News intensity based on aggregate, Arabic and Russian-language news has a significant negative impact on the return volatility in the next period, alt-hough the R-squared value is small.

The OOS forecasting performance of news intensity for all countries and languages is summarized in Table 7. It reports estimates of the coefficient (b_1) of the forecasts based on news intensity in Equation (4). News intensities have significant impacts on ex-post volatility. Many estimates are significant (in boldface). The majority of these cases indicates a negative impact of news intensity on volatility in the next period (in red).

To compare relative information contents of news intensity, historical volatility and the EGARCH model in OOS forecasting analysis, we consider forecasts of volatility based on these three alternatives simultaneously. Using the UK as an example, Table 8 shows the results from the encompassing test (Equation (5)) in the OOS analysis. First, consider two forecasts based on news intensities and the EGARCH model simultaneously. The EGARCH forecast has significant predictive power for ex-post volatility across all languages. News intensities shows significant predictive power in addition to the informational content in the EGARCH forecast for all languages except German. The direction of the impact is negative. The higher the news intensities, the lower the volatility in the next period. Once the forecast based on the historical volatility is

Table 7. Out-of-sample predictive power of news intensities

	Aggregate	Arabic	English	French	German	Russian
Australia	0.385	0.494	0.178	0.416	0.669	0.688
Austria	-0.101	-0.839	0.242	-0.349	0.069	-0.476
Belgium	0.117	-0.673	-0.025	-0.017	**-0.426**	-0.073
Brazil	**-1.599**	**-0.890**	**-1.556**	**-0.393**	-0.359	**-1.818**
Chile	-0.479	0.035	-0.285	-0.371	0.104	-0.164
China	**1.069**	-0.622	**1.126**	**1.181**	0.139	**1.029**
Colombia	**3.250**	**2.922**	**3.347**	**2.728**	**2.025**	**3.511**
Denmark	**-0.316**	**-0.667**	**-0.588**	-0.252	-0.258	-0.427
Finland	-0.048	-0.222	-0.069	-0.091	0.002	0.023
France	**-0.175**	-0.349	**-0.160**	-0.169	-0.134	-0.135
Germany	-0.211	-0.249	-0.272	-0.183	-0.222	-0.223
HongKong	-1.121	**-1.243**	-0.393	**-1.319**	**-2.777**	-0.512
India	**-2.749**	-0.264	**-2.973**	**-2.455**	-0.183	**-1.392**
Indonesia	**-1.228**	**-1.175**	**-1.143**	0.029	-0.546	**-1.795**
Ireland	**-0.074**	0.120	**-0.091**	0.043	**-0.124**	0.037
Italy	-0.125	-0.042	-0.141	-0.060	-0.095	-0.088
Japan	**-0.845**	**-0.444**	**-0.486**	**-1.020**	**-1.038**	**-0.861**
Malaysia	**3.309**	1.829	**3.064**	**2.884**	**3.556**	**3.589**
Mexico	-0.236	-0.088	-0.442	-0.130	**-1.916**	-0.103
Netherlands	**-0.220**	-0.133	**-0.288**	-0.115	-0.098	-0.168
New Zealand	**-0.838**	**-2.884**	**-0.991**	-0.130	**-1.334**	**-0.824**
Norway	-0.133	-0.741	-0.052	-0.203	-0.332	**-0.454**
Philippine	**0.840**	**-1.388**	**0.860**	0.271	**-2.643**	0.045
Poland	-0.231	**0.708**	-0.527	-0.268	-0.321	-0.635
Singapore	1.944	1.459	**2.208**	1.270	1.241	0.282
South Africa	**-0.933**	0.101	**-1.456**	-0.189	-0.121	**-0.954**
South Korea	1.013	**-1.678**	**0.915**	0.802	2.077	0.399
Spain	**-0.254**	**-0.297**	**-0.234**	**-0.119**	0.081	**-0.220**
Sweden	**-0.125**	-0.678	**-0.126**	**-0.139**	-0.014	-0.095
Switzerland	0.370	-0.165	0.323	0.213	0.259	0.205
Taiwan	0.162	-1.751	0.339	-0.448	-5.722	-0.718
Thailand	**-0.345**	-0.038	-0.253	**-0.734**	**-0.560**	-0.154
Turkey	-0.286	0.123	**-0.649**	-0.209	0.078	**-0.460**
UK	**-0.136**	**-0.962**	-0.160	-0.135	-0.093	**-0.192**
US	1.437	**2.000**	1.436	**1.524**	1.317	1.055

Note: Significant coefficients according to the t-test at 10% are in boldface. The relevant t-statistics use the White (1980) corrected standard errors. In addition, for these significant coefficients, they are in green if positive and in red if negative.

Table 8. Relative information content in the out-of-sample forecasting for the UK

$\sigma^2_{t+1} = b_0 + b_1\sigma^2_{St} + b_2\sigma^2_{Et} + b_3\sigma^2_{Ht} + \epsilon_{t+1}$

b_0	b_1	b_2	b_3	R^2
		Aggregate		
0.057	**-0.218**	**1.193**		0.049
0.054	-0.127	**0.684**	**0.367**	0.174
		Arabic		
0.355	**-1.248**	**1.284**		0.073
0.274	**-0.895**	**0.780**	**0.354**	0.188
		English		
0.072	**-0.272**	**1.218**		0.051
0.061	-0.151	**0.697**	**0.366**	0.175
		French		
0.062	**-0.231**	**1.199**		0.049
0.059	-0.143	**0.691**	**0.367**	0.175
		German		
0.044	-0.137	**1.123**		0.042
0.042	-0.061	**0.633**	**0.372**	0.172
		Russian		
0.079	**-0.269**	**1.207**		0.054
0.070	**-0.169**	**0.701**	**0.364**	0.177

Note: Significant coefficients according to the t-test at 10% are in boldface. The relevant t-statistics use the White (1980) corrected standard errors.

added, news intensities show a significant negative impact using Arabic and Russian.

The OOS predictive power of news intensities for the return volatility can be confirmed for many countries across all languages. Table 9 summarizes estimates of the coefficients (b_1) for the forecasts based on news intensities (σ^2_{St}) in a predictive regression with the forecasts based on the EGARCH model, and in a predictive regression with forecasts based on both the EGARCH model and on historical volatility. Many of the coefficients for news intensities are significant (in boldface), and the impact of news intensities on ex-post volatility is overwhelmingly negative: the red colour is dominating the green colour.

5 Discussion and conclusion

Our study examines, for the first time, how news coverage in different languages impacts stock markets around the world. To do so, we focus on how the volume of relevant news in each of five languages (as well as in all languages in aggregate) con-

Table 9. Relative information content of news intensities in the out-of-sample forecasting

	$\sigma^2_{t+1} = b_0 + b_1\sigma^2_{St} + b_2\sigma^2_{Et} + \epsilon_{t+1}$						$\sigma^2_{t+1} = b_0 + b_1\sigma^2_{St} + b_2\sigma^2_{Et} + b_3\sigma^2_{Ht} + \epsilon_{t+1}$					
	Aggregate	**Arabic**	**English**	**French**	**German**	**Russian**	**Aggregate**	**Arabic**	**English**	**French**	**German**	**Russian**
Australia	0.246	0.384	0.017	0.248	0.548	0.601	**0.823**	0.644	0.552	**1.135**	**1.332**	**1.210**
Austria	-0.245	**-1.573**	0.046	-0.396	-0.064	**-0.654**	-0.161	-1.276	0.076	-0.257	0.061	-0.480
Belgium	-0.234	**-1.725**	-0.320	-0.553	**-0.724**	-0.734	0.195	-0.129	0.068	0.356	-0.067	0.336
Brazil	**-1.231**	**-0.751**	**-1.274**	-0.198	-0.056	-1.116	**-1.198**	**-0.717**	**-1.250**	-0.185	-0.022	-1.032
Chile	-1.059	-0.493	-0.792	-0.976	-0.242	-0.555	-1.014	-0.497	-0.742	-0.929	-0.209	-0.518
China	0.659	-0.892	**0.711**	0.598	-0.271	0.551	0.632	-0.968	0.684	0.557	-0.336	0.536
Colombia	0.950	0.079	1.333	1.469	0.562	1.095	1.809	0.760	2.242	**2.139**	1.030	2.041
Denmark	**-0.351**	**-0.794**	**-0.659**	-0.277	-0.339	-0.490	**-0.370**	**-0.835**	**-0.702**	**-0.296**	-0.361	-0.517
Finland	**-0.196**	-0.801	**-0.231**	**-0.204**	-0.053	-0.129	**-0.170**	-0.677	**-0.201**	**-0.184**	-0.044	-0.102
France	**-0.237**	-0.307	**-0.204**	**-0.233**	-0.185	-0.197	**-0.203**	-0.332	**-0.174**	**-0.189**	-0.176	-0.158
Germany	**-0.362**	-0.238	**-0.445**	**-0.415**	**-0.379**	-0.280	**-0.319**	-0.334	**-0.387**	**-0.359**	-0.339	-0.254
HongKong	-0.983	-0.982	-0.422	-1.180	-2.448	-0.425	-1.009	-1.157	-0.417	-1.219	-2.582	-0.438
India	**-2.960**	-0.643	**-3.225**	**-2.654**	-0.456	**-1.207**	**-2.998**	-0.637	**-3.265**	**-2.685**	-0.462	**-1.215**
Indonesia	**-1.306**	**-1.078**	**-1.203**	-0.193	-0.483	**-1.733**	**-1.177**	**-1.008**	**-1.134**	-0.022	-0.465	**-1.522**
Ireland	**-0.075**	0.043	**-0.084**	0.010	**-0.133**	0.010	**-0.077**	0.016	**-0.085**	-0.002	**-0.133**	-0.000
Italy	**-0.283**	**-0.291**	**-0.306**	-0.102	**-0.182**	**-0.247**	-0.182	-0.174	-0.200	-0.058	-0.102	-0.145
Japan	**-0.787**	**-0.463**	**-0.480**	**-0.893**	**-0.919**	**-0.770**	**-0.795**	**-0.463**	**-0.482**	**-0.894**	**-0.921**	**-0.772**
Malaysia	1.750	0.290	1.661	**1.678**	1.342	2.272	**2.440**	0.971	**2.297**	**2.179**	2.690	**3.218**
Mexico	-0.253	-0.069	-0.436	-0.172	**-1.911**	-0.136	-0.318	-0.106	-0.509	-0.250	**-2.007**	-0.162
Netherlands	**-0.418**	-0.357	**-0.468**	**-0.286**	-0.189	**-0.356**	**-0.275**	-0.247	**-0.303**	-0.158	-0.158	-0.235
New Zealand	**-0.841**	**-2.899**	**-1.001**	-0.124	**-1.332**	**-0.824**	**-0.765**	**-2.723**	**-0.932**	-0.114	**-1.181**	**-0.738**
Norway	-0.318	-1.310	-0.218	-0.311	-0.582	-0.785	-0.253	-1.078	-0.161	-0.218	-0.390	-0.647
Philippine	**0.734**	**-1.200**	**0.759**	0.156	**-2.437**	-0.061	**0.735**	**-1.177**	**0.751**	0.179	**-2.408**	-0.050
Poland	-0.218	**0.886**	-0.521	-0.272	-0.304	-0.640	-0.222	0.876	-0.525	-0.264	-0.296	-0.639
Singapore	0.836	1.108	1.131	-0.900	-0.043	-2.025	1.309	1.019	1.679	-0.123	0.622	-1.192
South Africa	**-1.391**	-0.363	**-1.912**	**-0.553**	-0.273	**-1.440**	**-1.452**	-0.360	**-1.997**	**-0.581**	-0.282	**-1.536**
South Korea	-0.180	**-3.277**	-0.278	-2.161	0.688	**-1.640**	-0.177	**-3.366**	-0.277	-2.301	0.707	**-1.712**
Spain	**-0.413**	**-0.401**	**-0.387**	**-0.216**	0.040	**-0.417**	**-0.294**	**-0.318**	**-0.256**	**-0.146**	0.038	**-0.271**
Sweden	**-0.141**	**-1.450**	-0.128	**-0.297**	0.024	-0.138	**-0.114**	-1.197	-0.110	**-0.217**	0.017	-0.114
Switzerland	-0.161	-1.518	-0.186	-0.361	-0.244	-0.434	-0.160	-1.538	-0.185	-0.364	-0.244	-0.439
Taiwan	0.114	-1.762	0.294	-0.543	-5.976	-0.799	0.110	-1.774	0.300	-0.482	-5.594	-0.723
Thailand	**-0.324**	-0.040	-0.257	**-0.634**	**-0.464**	-0.203	-0.266	-0.050	-0.217	**-0.587**	**-0.483**	-0.120
Turkey	-0.283	0.130	**-0.649**	-0.204	0.086	**-0.458**	-0.237	0.082	**-0.559**	-0.176	0.072	**-0.378**
UK	**-0.218**	**-1.248**	**-0.272**	**-0.231**	-0.137	**-0.269**	-0.127	**-0.895**	-0.151	-0.143	-0.061	**-0.169**
US	-0.298	0.426	-0.353	0.064	-0.397	-0.644	-0.399	0.356	-0.471	-0.000	-0.454	-0.736

Note: Significant coefficients according to the t-test at 10% are in boldface. The relevant t-statistics use the White (1980) corrected standard errors. In addition, for these significant coefficients, they are in green if positive and in red if negative.

tributes to the conditional variance of returns for each of 35 countries' stock markets. This contribution is consistently significant - for example, in Table 9, which summarizes our findings, the proportion of coefficients significant at the 10% level is 37% - much higher than would be expected by chance. This suggests that news coverage indeed is an important variable to consider in modelling stock market volatility. Interestingly, greater news coverage is generally associated with lower subsequent volatility. A possible interpretation of this novel result is that an increased volume of country-related news is associated with resolution of uncertainty with regard to that country. Related to this, a richer information environment may have a reassuring effect on investors - or conversely, a lack of news coverage could make investors jittery.

Our results by language, however, are harder to interpret. One could have expected English-language news to be the dominant driver of any effect of news coverage on volatility across the globe. This is only occasionally the case. Further, we would expect each country's news in its official language to have a significant impact on the volatility of that country's stock market. While this is the case for France and French-language news, it is not so for Belgium and French-language news, nor (in our out-of-sample tests) for Germany and German-language news. We expect to be able to shed more light on this in the next version of the chapter, when we will add Chinese, Danish and Spanish-language articles to our analyses, as well as extend our sample period by an additional seven months to May 2019.

While our research embraces the "big data" dimension of news analytics, the sheer volume of analysed text presents challenges that translate into limitations for the current study. First, our news data are scraped daily from tens of thousands of online news websites. In calculating news intensities, we weight all of these websites equally. A more sophisticated weighting scheme may lead to easier-to-interpret results. Further, in order to process vast quantities of textual data at a manageable cost in terms of computing power, we needed a suitably efficient text-analytic implementation. Although semantic fingerprinting can cope with the volume of data, the flip side of this efficiency is relatively low granularity of the semantic space, adding noise to our news intensity variables. Despite these limitations, our analyses have shown, for the first time, that all-encompassing, multilingual measures of related news coverage do impact asset volatility across a wide range of countries. Drilling down into the news coverage data and increasing the granularity of the text analytics, as well examining the behaviour of other asset classes such as individual stocks, industry indices/ETFs and commodities are promising directions for future research.

References

[1] Bali, Turan, Andriy Bodnaruk, Anna Scherbina, and Yi Tang, 2017, Unusual news events and the cross-section of stock returns, Management Science, forthcoming.

[2] Baloria, V.P. and Heese, J., 2018. The effects of media slant on firm behaviour, *Journal of Financial Economics* 129, 184-202.

[3] Barber, Brad M. and Terrance Odean, 2008, All that glitters: The effect of attention and news on the buying behaviour of individual and institutional investors, *Review of Financial Studies* 21, 785–818.

[4] Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* 31, 307–327.

[5] Boudoukh, Jacob, Ronen Feldman, Shimon Kogan and Matthew Richardson, 2018, Information, trading, and volatility: Evidence from firm-specific news, forthcoming, *Review of Financial Stud-ies*.

[6] Boukus, Ellyn, and Joshua V. Rosenberg, 2006, The information content of FOMC minutes, Work-ing Paper, *Federal Reserve Bank of New York*.

[7] Box, Travis, 2018, Qualitative similarity and stock price comovement, *Journal of Banking and Finance* 91, 49-69.

[8] Byström, Hans, 2016, Language, news and volatility, *Journal of International Financial Markets*, Institutions and Money 42, 139-154.

[9] Calomiris, Charles W. and Harry Mamaysky, Harry, 2018, How news and its context drive risk and returns around the world, *Journal of Financial Economics*, forthcoming.

[10] Cohen, L., Malloy, C. and Nguyen, Q., 2018. Lazy prices (No. w25084). *National Bureau of Eco-nomic Research*.

[11] Da, Zhi, Joseph E. Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461-1499.

[12] Day, Theodore E. and Craig M. Lewis, 1992, Stock market volatility and the information content of stock index options, *Journal of Econometrics* 52, 267-287.

[13] Dimson, Elroy, Paul Marsh and Mike Staunton, 2018, *Credit Suisse Global Investment Returns Yearbook*.

[14] Engelberg, Joseph E. and Christopher A. Parsons, 2011, The causal impact of media in financial markets, *Journal of Finance* 66, 67-97.

[15] Fang, Lily H. and Joel Peress, 2009, Media coverage and the cross-section of stock returns, *Journal of Finance* 64, 2023-2052.

[16] Fedyk, Anastassia and James Hodson, 2017, When can the market identify old news?, Working Paper, Harvard University.

[17] Glasserman, Paul and Harry Mamaysky, 2018, Does unusual news forecast market stress?, *Journal of Financial and Quantitative Analysis*, forthcoming.

[18] Glosten, Lawrence R., Ravi Jagannathan, David E. Runkle, 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* 48, 1779-1801.

[19] Heston, Steven L., and Nitish Ranjan Sinha, 2017, News versus sentiment: Predicting stock re-turns from news stories, *Financial Analysts Journal* 73, 67-83.

[20] Huberman, Gur and Tomer Regev, 2001, Contagious speculation and a cure for cancer: A non-event that made stock prices soar, *Journal of Finance* 56, 387-396.

[21] Ibriyamova, Feriha, Samuel Kogan, Galla Salganik-Shoshan and David Stolin, 2017, Using semantic fingerprinting in finance, *Applied Economics* 49, 2719-2735.

[22] Ibriyamova, Feriha, Samuel Kogan, Galla Salganik-Shoshan and David Stolin, 2019, Predicting stock return correlations with brief company descriptions, *Applied Economics* 51, 88-102.

[23] Jiang, Hao, Sophia Zhengzi Li and Hao Wang, 2017, News Momentum, Working Paper, Michigan State University.

[24] Larsen, Vegard H. and Leif Anders Thorsrud, 2015, The value of news, Working Paper, BI Norwe-gian Business School.

[25] Loughran, Tim and Bill McDonald, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187-1230.

[26] Nelson, D. B., 1991, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica* 59(2), 347–370. Shiller, Robert J., 2017, Narrative economics, *American Economic Review* 107, 967–1004.

[27] Solomon, D. H., 2012, Selective publicity and stock prices, *Journal of Finance* 67, 599–638.

[28] Tao, Ran, Christ Brooks and Adrian Bell, 2019, Tomorrow's fish and chip paper? Slowly incorpo-rated news and the cross-section of stock returns, Working Paper, ICMA Center.

[29] Taylor, Stephen J., 1986, Modelling Financial Time Series. *Wiley*, Chichester.

[30] Tetlock, Paul, 2010, All the news that's fit to reprint: Do investors react to stale information?, *Review of Financial Studies* 24, 1481–1512.

[31] Vozlyublennaia, N., 2014. Investor attention, index performance, and return predictability, *Journal of Banking and Finance* 41, 17–35. Webber, Francisco Eduardo de Sousa, 2015, Semantic folding and its application in semantic fingerprinting, Cortical.io White Paper, Version 1.0, available from http://www.cortical.io/static/downloads/semantic-folding-theory-white-paper.pdf

[32] White, Halbert, 1980, A heteroscedasticity-consistent covariance matrix estimato and a direct test for heteroscedasticity, *Econometrica* 48, 817-838.

[33] Yermack, David, The Michelle markup: The first lady's impact on stock prices of fashion compa-nies, 2010, Working Paper, New York University.

A Semantic fingerprint of the news item and its Sample Webhose.io news item

<?xml version="1.0" encoding="UTF-8"?>
</document>
<type>mainstream</type>
<forum>http://www.moneycontrol.com/rss/latestnews.xml</ forum >
<forum_title>Moneycontrol Latest News</forum_title>
<discussion_title>RBI 'not a cheerleader', but it still cheers markets in 2015</discussion_title>
<language>english</language>
<gmt_offset/>
<topic_url>http://www.moneycontrol.com/news/economy/rbi-notcheerleaderit-still-cheers-markets2015_4813601.html</ topic_url >
<topic_text> Dec 31, 2015, 02.20 PM | Source: PTI RBI 'not a cheerleader', but it still cheers markets in 2015 Raghuram Rajan's term ends on September 3, 2016 and whether the outspoken Governor gets an extension, like his most of his predecessors, or not is something that will be keenly watched in the New Year. Like this story, share it with millions of investors on M3 RBI not a cheerleader, but it still cheers markets in 2015 Raghuram Rajan's term ends on September 3, 2016 and whether the outspoken Governor gets an extension, like his most of his predecessors, or not is something that will be keenly watched in the New Year. Post Your Comments Share Cancel He always ruled out being a 'cheerleader for the markets', but it was the multiple rate cuts by RBI Governor Raghuram Rajan and his ceaseless efforts to keep inflation under check that turned out to be the biggest cheers for the Indian financial markets in 2015. It was one of his most famous oneliners -- "I am Raghuram Rajan and I do what I do" -- that eventually summed up the year 2015 for the central bank and the same is expected to hold true in the New Year as well when RBI cracks its whip on banks to clean up their balancesheets and also compete with the financial markets when it comes to being source for funds. Raghuram Rajan's term ends on September 3, 2016 and whether the outspoken Governor gets an extension, like his most of his predecessors, or not is something that will be keenly watched in the New Year. It would be interesting in the context of the public postures he has taken on many issues ranging from 'Making in India' to his reference to Hitler and on intolerance, which have been interpreted in various quarters as being against the current regime. At the same time, there have been words doing the rounds that Rajan is personally liked by the Prime Minister and there have been rumours about he being in race for some larger roles including the next head of International Monetary Fund, where he has served in the past as the Chief Economist. Looking back at 2015, the year will also be remembered for cutting short the wider powers of the central bank, including the Governor's own prerogative to set the rates or moving the public debt management out of the Mint Road. During the year, RBI became an inflation-targeting central bank, while Rajan also kick-started the era of differentiated banking by giving out in-principle approvals to 11 payments banks and 10 small finance banks. While the aspirants of small finance banks are populated largely by microfinance institutions, the payments bank licencees are storeyed names from the corporate world such as Ambanis, Birlas, Mahindras as also the telecom biggies like Airtel and Vodafone. With certainty around achievement of the near-term inflation objective of 6 percent increasing, Rajan made a dramatic shift in his policy stance early in the year towards being accommodative and announced a surprise rate cut on January 15. He followed it up with three similar moves during the year, cutting rates by a cumulative 1.25 percent in 2015. All his rate cuts were welcomed by the stock markets with huge rallies, while inflation remaining under check gave the government much-required legroom in its policy moves. The 6-percent headline inflation target for January 2016 is part of the 'glide path' suggested by the deputy governor Urijit Patel-led Committee, and for RBI the tougher task of yanking it down to 4 percent in next two years starts now. </topic_text>
<spam_score>0.00</spam_score>
<post_num>1</post_num>
<post_id>post-1</post_id>
<post_url>http://www.moneycontrol.com/news/economy/rbi-notcheerleaderit-still-cheers-markets2015_4813601.html</post_url >
<post_date>20151231</post_date>
<post_time>0847</post_time>
<external_links> </external_links>
<country>IN</country>
<main_image>http://www.moneycontrol.com/news_image_files/2014/356x200/r/RBI-New_16-9_356x200_3005_356.jpg</main_image>
</document>

Abstract

My thesis comprises three chapters that revolve around the application of textual data analysis to explore various aspects of finance and industrial organization. In Chapter 1, I delve into the development of a dynamic global text-based industry classification that outperforms conventional industry classifications by generating more coherent groups of firms, particularly in emerging economies. Chapter 2 investigates the significance of product relatedness in mergers and acquisitions on the global market. To measure pairwise product similarity, I employ textual analysis techniques. Lastly, Chapter 3 explores the influence of news from different countries and languages on stock market indices. By employing a comprehensive dataset of news articles and employing multi-lingual text processing techniques, I aim to evaluate the impact of news in elucidating market volatility.

Keywords: Textual Analysis, Industry Classification, Mergers and Acquisitions, Product Similarity, News Impact on Stock Markets, Multi-Lingual Text Processing, Asset Complementarity

Résumé

Ma thèse comprend trois chapitres qui gravitent autour de l'application de l'analyse de données textuelles pour explorer différents aspects de la finance et de l'organisation industrielle. Dans le chapitre 1, j'approfondis le développement d'une classification industrielle mondiale dynamique basée sur le texte, qui surpasse les classifications industrielles traditionnelles en générant des groupes d'entreprises plus cohérents, notamment dans les économies émergentes. Le chapitre 2 examine l'importance de la similarité des produits dans les fusions et acquisitions sur le marché mondial. Pour mesurer la similarité des produits deux à deux, j'utilise des techniques d'analyse textuelle. Enfin, le chapitre 3 explore l'influence des nouvelles provenant de différents pays et langues sur les indices boursiers. En utilisant un ensemble de données exhaustif d'articles de presse et des techniques de traitement de texte multilingues, je vise à évaluer l'impact des nouvelles dans l'explication de la volatilité du marché.

Mots clés: Analyse Textuelle, Classification Industrielle, Fusions et Acquisitions, Similarité de Produits, Impact des Nouvelles sur les Marchés Boursiers, Traitement de Texte Multilingue, Complémentarité des Actifs