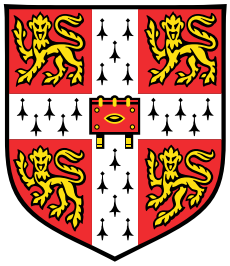
Essays on Probabilistic Machine Learning for Economics



Nikolas Kuhlen

Faculty of Economics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College July 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text and Acknowledgements. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 60,000 words.

July 6, 2021

Nikolas Kuhlen

Acknowledgements

My doctoral journey certainly was not the most conventional and has benefited from the support of many people. First and foremost, I want to thank my supervisor Vasco Carvalho and my mentor and collaborator Mirko Draca. This thesis would not have been possible without their guidance and support. I thank Vasco for pushing me to always strive for deeper insights while keeping in mind the bigger picture. Mirko was extremely generous with his time and our uncountable Skype sessions were invaluable for me to learn how to let the data speak and ask meaningful questions.

I thank my squash partner Andrew Preston for being a great co-author and sharing his knowledge on macroeconomics with me. The amazingly smooth and stimulating collaboration with him resulted in the first chapter of this thesis.

I am grateful to Mingli Chen for her support and creating many opportunities for me to grow as a researcher. I thank Stephen Hansen for introducing me to the mixed-membership models literature during my weekly trips to Oxford in my first year which was one of the main influences for the eventual research path I took. I am thankful to Chenlei Leng for allowing me to switch from Statistics to Economics at the end of my second year. This switch would also not have been possible without Louise Cross, Amy Gallimore, Joanna Gathercole, Rohan Lourdelet, Ben Murton, and Samantha Selvarajah.

The Turing provided a fantastic environment for me to carry out my interdisciplinary research. I was fortunate enough to be part of a wonderful doctoral cohort and the wider community at the Turing. In particular, I thank my friends Andrea Pizzoferrato and Prateek Gupta for the great discussions and fun we had inside and outside of the office. I would like to thank my advisor Christopher Rauh and Bill Janeway at the Faculty of Economics in Cambridge for their help and valuable suggestions. Thank you also to my colleagues at the Big Data Lab at HSBC for teaching me how to work with genuinely large data sets.

I am indebted to my family Sabine, Thomas and Sebastian for being a source of unconditional support at all times. Lastly, I thank Yonglin for her continuous help and emotional support in dealing with all the ups and downs of this journey.

Contents

List of Figures

List of Tables

Introduction

> *"You should call it entropy, for two reasons. In the first place, the formula has been used in statistical mechanics under that name. In the second place, and more importantly, no one knows what entropy really is, so in a debate you will always have the advantage."*
>
> — John von Neumann to Claude Shannon.

This thesis consists of three essays that explore the use of probabilistic machine learning techniques in combination with information-theoretic concepts to answer economic questions. Over the past years, economists have started applying machine learning methods to a wide range of topics. Probabilistic methods in the context of unsupervised learning represent one particular modelling approach at the intersection of computer science and statistics. While widely used in applied statistics, these models, however, do not necessarily provide relevant and interpretable outputs from an economist's perspective. In this thesis, I appeal to information-theoretic methods to summarise the probabilistic information inferred from such models and construct economically meaningful measures.

Specifically, I employ a combined framework that builds on the family of mixed-membership models. Mixed-membership models have emerged over the past two decades as a flexible classification-like modelling tool for the unsupervised analysis of high-dimensional multivariate data. In contrast to other approaches where each observational unit belongs to a single category or cluster, the underlying generative process assumes that every unit partially belongs to all clusters. This information is captured by an individual distribution over clusters expressed in terms of a membership vector (Airoldi, Blei, Erosheva, and Fienberg, 2014).

While the shared membership of units across categories is central to the definition of mixed-membership models, few analyses make full use of this feature. In particular, most studies focus on identifying and interpreting extreme, ideal, or edge types. That is, they discard the information provided by the mixed-membership representation and instead use the models for crisp clustering (Singer and Castro, 2014).

From the perspective of information theory, each mixed-membership distribution in the generative process represents a source that produces a signal – the observed outcomes for each unit. The following chapters build on this interpretation and compute information-theoretic quantities to describe the characteristics of the mixed-membership distributions – thereby taking into account all of the available information about an observational unit.

A small recent literature outside of economics has used this combination of mixed-membership models – more specifically topic models – and information-theoretic measures across different research areas such as the cognitive sciences (Murdock, Allen, and Dedeo, 2017), history of political thought (Barron, Huang, Spang, and DeDeo, 2018), and cultural evolution (Jing, DeDeo, and Ahn, 2019). The methodological contribution of this thesis is to adapt, apply and extend this framework to the economic domain as follows.

Chapter 1: News Entropy

Chapter 1, which is joint work with Andrew Preston, introduces the concept of 'news entropy' to characterise the relationship between news coverage and the economy. Intuitively, news entropy decreases as the news focus on a smaller set of pressing topics. We observe that news entropy exhibits clear negative spikes close to important economic, financial, and political events. Investigating the effect of changes in news entropy, we find that decreases are associated with two key features: an increase in uncertainty measures and a macroeconomic contraction. The variable is priced in the cross-section of stock returns and low news entropy is associated with increased stock price volatility at the firm level.

Chapter 2: Exploration and Exploitation in US Technological Change

Chapter 2, which is joint work with Vasco M. Carvalho and Mirko Draca, investigates the question: How do firms and inventors move through 'knowledge space' as they develop their innovations? We propose a method for tracking patterns of 'exploration and exploitation' in patenting behaviour in the US for the period since 1920. Our exploration measure is constructed from patent texts and involves the use of 'Bayesian Surprise' to measure how different current patent-based innovations are from existing portfolios. Our results indicate that there are distinct 'life-cycle' patterns to firm and inventor exploration. Furthermore, exploration activity is more geographically concentrated than general patenting, but this concentration is centred outside the main hubs of patenting.

Chapter 3: Endogenous Technology Space

Chapter 3 spans a new endogenous technology space from patent texts. I then rely on information-theoretic methods to construct measures of technological firm distances – both fixed and time-varying. Using the latter, I present three sets of findings. First, I observe that industries are becoming more technologically specialised and segregated over time. Second, I identify the emergence of internet companies in the mid-1990s as a distinct group of firms with roots in traditional information and communication technologies. Third, I determine the unique set of time-varying rivals surrounding a focal firm in the endogenous technology space. We demonstrate the validity of this approach by means of a case study of the software company Oracle.

References

Airoldi, Edoardo M., David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg. 2014. *Handbook of mixed membership models and their applications*, 1–572. DOI: 10.1201/b17520. [1]

Barron, Alexander T.J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, institutions, and innovation in the debates of the French Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (18): 4607–12. DOI: 10.1073/pnas.1717729115. arXiv: 1710.06867. [2]

Jing, Elise, Simon DeDeo, and Yong-Yeol Ahn. 2019. "Sameness Attracts, Novelty Disturbs, but Outliers Flourish in Fanfiction Online." arXiv: 1904.07741. URL: http://arxiv.org/abs/1904.07741. [2]

Murdock, Jaimie, Colin Allen, and Simon Dedeo. 2017. "Exploration and exploitation of Victorian science in Darwin's reading notebooks." *Cognition* 159: 117–26. DOI: 10.1016/j.cognition.2016.11.012. [2]

Singer, Burton H., and Marcia C. Castro. 2014. "Interpretability constraints and trade-offs in using mixed membership models." In *Handbook of Mixed Membership Models and Their Applications*, 159–72. DOI: 10.1201/b17520. [1]

Chapter 1

News Entropy*

Joint with Andrew Preston

1.1 Introduction

The state of the economy is an integral factor that shapes news coverage. During times of economic and financial crises as well as political uncertainty, the news are likely to cover a smaller set of pressing topics, becoming more concentrated than in normal times. Our understanding of this relationship, however, is limited. In particular, the concept of information in the news is difficult to both conceptualise and measure empirically. Previous research in this area has typically focused on connecting specific terms, topics or sentiments to aggregate economic indicators. This approach, however, is prone to bias, arbitrary linguistic choices and usually suffers from limited generalisability due to underlying changes and differences in languages.

In this paper, we introduce the concept of 'news entropy' to characterise the relationship between the news and the economy. In particular, we first quantify the information communicated by newspaper articles. In doing so we build on an information-theoretic approach to statistical natural language processing. This yields a well-defined measure of news entropy with a number of desirable properties. The underlying intuition is as follows. If the news focus on a small number of topics, news entropy is low. Conversely, news entropy is high in times when the news cover a larger set of topics. In this sense, news entropy can be interpreted as capturing the degree of heterogeneity of news coverage and is related to the newsworthiness of current events.

* We thank Elliott Ash, Vasco M. Carvalho, Simon DeDeo, Milena Djourelova, Mirko Draca, Kristoffer Nimark, and Max Winkler for helpful comments. We thank participants at the UCLA-Warwick Machine Learning Seminar and the Economics and Data Science Seminar at ETH Zurich for useful feedback and suggestions. Kuhlen gratefully acknowledges the financial support of The Alan Turing Institute under research award No. TU/C/000030. Preston gratefully acknowledges the financial support of the Economic and Social Research Council.

Estimating the monthly news entropy for full texts of Wall Street Journal articles between 1984 and 2017, we observe that news entropy exhibits clear negative spikes during economic events such as the financial crises in 2008 and 2012, political events, and close to presidential elections. We also find a strong negative correlation with widely used news-based measures such as newsworthiness and policy uncertainty indices. We then empirically investigate the effect of changes in news entropy with respect to the economy. Our results indicate that decreases in news entropy are associated with two key features: a rise in uncertainty and a macroeconomic contraction. Additionally, we demonstrate that news entropy is priced in the cross-section of stock returns, and that low entropy is associated with increased stock price volatility at the firm level.

More specifically, to measure news entropy, we first rely on topic distributions obtained from applying Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003) to the corpus of Wall Street Journal articles. We then define news entropy as the Shannon entropy (Shannon, 1948) of the monthly topic distributions. Thus, in contrast to other unitless indicators and indices, news entropy is measured in bits – a proper unit grounded in information theory. Note that due to this construction, news entropy is language-agnostic and thus highly generalisable. In addition to the overall entropy measure estimated for the entire set of news topics, we also estimate the entropy of thematically related subsets – namely cultural, economic and political news.

Examining the relationship between news entropy and other news-based measures, we find that political news entropy is strongly negatively correlated with the concept of news pressure by Eisensee and Strömberg (2007). At the same time, we observe no significant relationship of news pressure with economic news entropy, suggesting that top news stories on TV are dominated by political events. We also find that news entropy is negatively correlated with the Economic Policy Uncertainty Index by Baker, Bloom, and Davis (2016) implying that news entropy, while a much broader concept, captures part of the notion of policy uncertainty.

Following Baker, Bloom, and Davis (2016), we then examine the firm-level impact of news entropy using option-implied stock price volatility as a proxy for firm-level uncertainty. We find that firms with higher exposure to government purchases are likely to show increased stock price volatility during periods of low news entropy. Additionally, we observe that news entropy subsumes the effects of both the Economic Policy Index by Baker, Bloom, and Davis (2016) as well as the Chicago Board Options Exchange Volatility Index (VIX) when including all three in the regression specification.

Next, we estimate the macroeconomic impact of news entropy fluctuations by identifying shocks as changes to news entropy which are orthogonal to the state of the economy. Using the local projection method of Jordà (2005), we estimate the

impulse responses for a set of key macroeconomic variables and find that a fall in entropy leads to a persistent fall in output and a V-shaped decline in employment which is followed by a subsequent overshoot. The shock is followed instantaneously with a rise in several extant measures of uncertainty.

A third key finding is that news entropy, as well as the economic and political news entropy measures, are a priced risk factor in the cross-section of stock returns. Given that we find decreases in entropy precede periods of severe economic distress, this result aligns with the rare disasters asset pricing model of Barro (2006), Gabaix (2012) and Wachter (2013).

Related Literature

The paper relates to several strands of research. It perhaps most prominently connects to the recent economics literature applying topic models and specifically Latent Dirichlet Allocation by Blei, Ng, and Jordan (2003) to various text data sources. To our knowledge, Mahajan, Dey, and Haque (2008), Fligstein, Brundage, and Schultz (2014), and Hansen, McMahon, and Prat (2017) are the first uses of Latent Dirichlet Allocation in an economics context.

Within this literature, our work is part of a small collection of papers that uses topic models to connect news language to economic activity. For example, similar to our paper, Bybee, Kelly, Manela, and Xiu (2019), Larsen and Thorsrud (2019) and Rauh (2019) use Latent Dirichlet Allocation to analyse news texts. They focus on empirically identifying those topics with the highest predictive power for aggregate economic outcomes. In contrast, our approach is much broader as it does not focus on the shares of specific topics but rather captures structural and behavioural patterns in the news. Nimark and Pitschner (2019) combine empirical observations from topic models with a theoretical framework. They document empirically that two major events increased the homogeneity in coverage across different newspapers devoting more front page coverage to them than to any other topic. News entropy provides a direct measure for this phenomenon as evidenced by our results for Wall Street Journal newspaper articles. Moving beyond the standard topic model, Bertsch, Hull, and Zhang (2021) apply the dynamic embedded topic model to identify economic narratives from Swedish newspaper articles. Using within-topic entropy, they find that the consolidation of narratives is strongly, positively associated with GDP growth over the business cycle. Conversely, they observe that narratives tend to fragment into competing explanations during macroeconomic contractions.

Our focus on newspaper coverage also links our work to Nimark (2014), who illustrates how media coverage of certain events can have definitive business cycle implications. A central principle of the framework developed is that highly concentrated news coverage should cause agents to suffer from higher uncertainty which

then spills over detrimentally to output and inflation. This is precisely the result we find empirically, as our identification strategy attempts to separate the portion of news concentration which arises endogenously from the state of the economy. Chahrour, Nimark, and Pitschner (2019) develop a model in which sectoral news coverage can be a substantial contributor to business cycle fluctuations. In a similar vein, Peress (2014) uses newspaper strikes to identify the causal effect of newspaper coverage on financial markets, finding that on strike days, stock market volatility is significantly reduced relative to normal trading days. This would imply that newspaper coverage is a vital component of the propagation mechanism of uncertainty, aligning with our empirical results.

More generally, our paper also relates to the recent economics literature constructing various indices from news language. For instance, the Economic Policy Uncertainty index by Baker, Bloom, and Davis (2016) counts the occurrence of a small set of policy-relevant terms in newspaper texts to measure uncertainty. Another prominent example from the political economy literature is the concept of news pressure by Eisensee and Strömberg (2007) which measures the airtime of the top three segments in news broadcasting. Manela and Moreira (2017) also use machine learning techniques to analyse the content of newspapers, but focus specifically on gauging the perceived risk of a rare economic disaster. Moreover, their analysis only concentrates on the front page of newspapers, whereas our approach is broader.

The remainder is organised as follows. Section 1.2 introduces the methodology of our news entropy measure and discusses its properties. Section 1.3 estimates the news entropy series and presents our main descriptive results. Section 1.4 investigates the relationship between news entropy and the economy from a firm, macroeconomic, and financial perspective. Section 1.5 concludes.

1.2 Methodological Framework

This section introduces the methodology of our measure. Section 1.2.1 describes Latent Dirichlet Allocation. Section 1.2.2 introduces Shannon entropy. This is followed by the definition of our news entropy measure and a discussion of its properties in Section 1.2.3.

1.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a hierarchical Bayesian model for text data (Blei, Ng, and Jordan, 2003). LDA generates documents from distributions over topics.

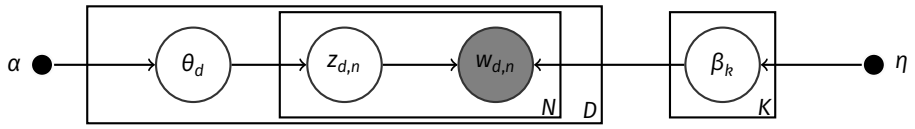


Figure 1.1. Latent Dirichlet Allocation.

Note: Shows the graphical model for Latent Dirichlet Allocation. Shaded variables are observed. Plates indicate replication of the nodes by the number in the lower right corner.

The topics are defined as probability vectors assigning a weight to each word in the vocabulary. That is, a topic is characterised by the set of words that it is most likely to use. Formally, LDA is specified in terms of the following process to generate a set of observed documents:

(1) For each document d:

   a. Draw topic proportions $\theta_d|\alpha \sim \text{Dir}(\alpha)$.

   b. For each word $w_{d,n}$:

      i. Draw assignment $z_{d,n}|\theta_d \sim \text{Mult}(\theta_d)$.

      ii. Draw word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.

where K specifies the number of topics, $\beta_{1:K}$ are the topic specific word distributions over the vocabulary, and α is a K-dimensional Dirichlet parameter. θ_d represents the topic proportions, z_d denotes the topic assignments, and w_d are the observed words for the d-th document. Figure 1.1 shows the corresponding graphical model.

To put this in words, each document is endowed with a Dirichlet-distributed vector that specifies the topic proportions. For each word in the document corpus, the model draws a topic assignment based on the topic proportions. Finally, the topic assignment is then used to generate the word. Note that this modelling approach implies that a word can be used for multiple topics with different probabilities. There is a variety of inference procedures for parameter estimation including sampling and optimisation based algorithms.

1.2.2 Shannon Entropy

The Shannon entropy (Shannon, 1948) of a random variable X is defined as

$$H(X) = -\sum_{i=1}^{N} p(x_i)\log_2 p(x_i),$$

where N is the number of possible outcomes and $p(x_i)$ is the probability of the outcome x_i. This can also be written as $H(\boldsymbol{p})$, where $\boldsymbol{p}$ is a vector of probabilities $(p_1, p_2, \ldots, p_N)$. When using a logarithm with base two, Shannon entropy is measured in bits.

There are many interpretations of entropy. From an information-theoretic perspective, entropy measures the amount of information that a random process carries about the outcome. It can also be interpreted as a measure of the uncertainty in a process. That is, it represents the uncertainty regarding the realisation of the random variable. In this paper, we rely on entropy to measure the degree of heterogeneity of a probability distribution. In particular, a decrease in entropy decreases the heterogeneity – or increases the homogeneity – of the random variable's outcomes.

Shannon entropy has the following properties. First, it is continuous with respect to the probabilities of the outcomes. Second, it is symmetric with respect to the order of the probabilities. Third, it is maximised when all probabilities $p(x_i)$ are equal. The maximum is equal to $\log_2(N)$. Fourth, the entropy of a process is equal to zero if all but one probability $p(x_i)$ are equal to zero. Fifth, if a process is divided up into successive processes, the original entropy is equal to the weighted sum of the individual entropies. We provide an interpretation of these properties in the context of our application in the following section.

1.2.3 News Entropy

Based on the definition of LDA and Shannon entropy, we now construct our measure of news entropy. From an information-theoretic perspective, each topic distribution in the generative model of LDA represents a source that produces a signal (Murdock, 2019). The signal is the stream of words forming the document. In this context, we define news entropy as the entropy of the topic distribution

$$\mu_d = H(\theta_d).$$

That is, μ_d represents the degree of heterogeneity of the outcomes of the process described by the topic distribution. Alternatively, news entropy can also be interpreted as a measure of uncertainty regarding the topic a word was generated from.[1]

The underlying intuition of news entropy is as follows. When an important event occurs, the news will dedicate a large share of their coverage to the event in question. In other times, when no major news event has occurred, the news instead cover several minor events. Assuming that different types of events can be represented by news topics, newspaper texts will be dominated by fewer topics during major events compared to normal times and secondary news are crowded out. In this sense, news entropy captures the degree of heterogeneity of news coverage and is related to the newsworthiness of current events.

1. When viewed as the uncertainty of the reader regarding the topic assignment the next word in the newspaper article, news entropy connects to the first use of entropy applied to natural language by Shannon (1951).

Further Properties

We now derive further properties of our measure. First, due to the continuity property of Shannon entropy, small changes in the topic shares result in small changes to the overall information. As a result, there are no discontinuous jumps in news entropy as the newspaper increases or decreases its focus on a particular topic.

Second, news entropy is invariant to changes in the ordering of topics inferred from LDA. This is a necessary condition for a proper measure both from a statistical and economic perspective. In particular, the ordering of topics might differ between different runs of the LDA algorithm on the same data. This is due to the random sampling as part of the computational inference procedure. As long as the inferred probabilities are the same, however, news entropy does not change. That is, we implicitly assume that the order in which the reader learns about the topics does not affect their information processing. Further, this implies that the topics' information shares are independent of each other. This independence assumption relates to the underlying assumption in the LDA model that topics are uncorrelated.[2]

Lastly, as stated above, entropy satisfies the following property: if a process is divided up into successive processes, the original entropy is equal to the weighted sum of the individual entropies. This can be interpreted as the "coarse-graining" property (DeDeo, Hawkins, Klingenstein, and Hitchcock, 2013). The coarse-graining property of entropy has three major implications for our application to topic distributions. First, specifying a larger number of topics in the LDA model results in higher news entropy since more information needs to be communicated. Second, the entropy of topic subsets after renormalising the topic shares will be smaller than the entropy for the whole set of topics. Third, for a given number of topics, we can coarse-grain the topic shares to calculate news entropy based on thematically related groups of topics. This is important since it allows to independently estimate the topic model with the number of topics set to be statistically optimal or provide the most intuitive interpretation of topics. This is then followed by calculating the entropy at the desired level of coarse-graining. This emphasises the flexibility and general applicability of our approach.

1.3 Estimation

This section estimates news entropy and presents our main descriptive results. Section 1.3.1 describes our original data sources. Section 1.3.2 describes the estimated

2. This is due to the independence assumption implicit to using Dirichlet distributed topic proportions. Under the Dirichlet, the topic shares are nearly independent. As a result, the presence of one topic is not correlated with the presence of another. To allow for a covariance structure between topics, Blei and Lafferty (2007) have developed the Correlated Topic Model.

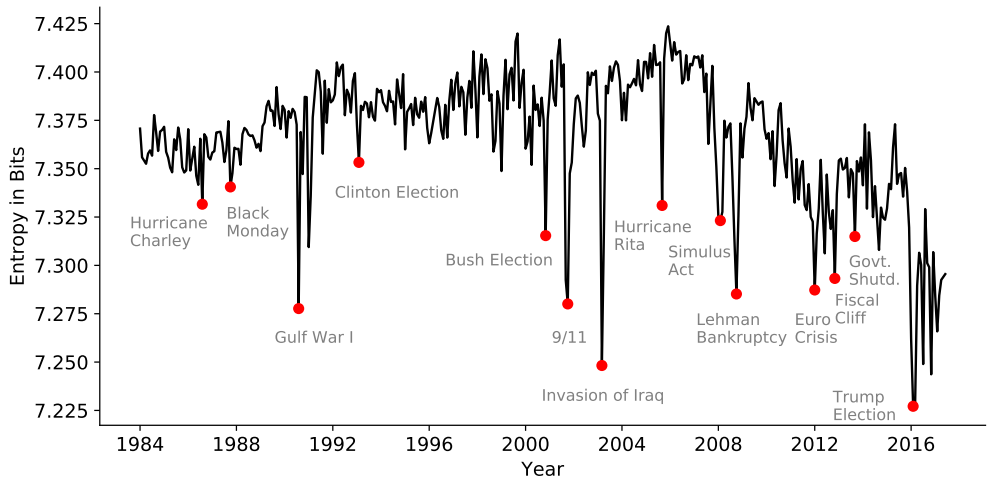


Figure 1.2. News Entropy.

Note: This figure shows news entropy from 1984 to 2017.

entropy series and connects them to major events. Section 1.3.3 compares news entropy to existing measures of news pressure and policy uncertainty.

1.3.1 Data

We rely on the pre-trained LDA topic vectors for the Wallstreet Journal (WSJ) provided by Bybee et al. (2019). The data set consists of the monthly topic vectors estimated from the full newspaper texts of 763,887 articles published between January 1984 and June 2017. The vocabulary comprises 18,432 uni-grams and bi-grams. For content consistency, articles published in sections other than the three core sections ("Section One," "Marketplace," and "Money and Investing") are excluded. In addition, articles with predominantly non-economic tags as well as regular data tables are excluded. The number of topics in the LDA model was set to 180 based on statistical goodness-of-fit criteria. Bybee et al. provide a data-driven hierarchy of increasingly broad meta-topics based on the semantic distances between topics. At the broadest level, the hierarchy distinguishes between "economy" topics and "politics and culture" topics. The macroeconomic data comes from the FRED-MD database of McCracken and Ng (2016). We obtain firm-level data from Baker, Bloom, and Davis (2016).

1.3.2 Descriptives

We calculate news entropy from the pre-trained WSJ topic distributions. Figure 1.2 shows the resulting time series from 1984 to 2017. The graph exhibits clear negative spikes during events related to the financial crises in 2008 and 2012, the Gulf and

Table 1.1. Correlations.

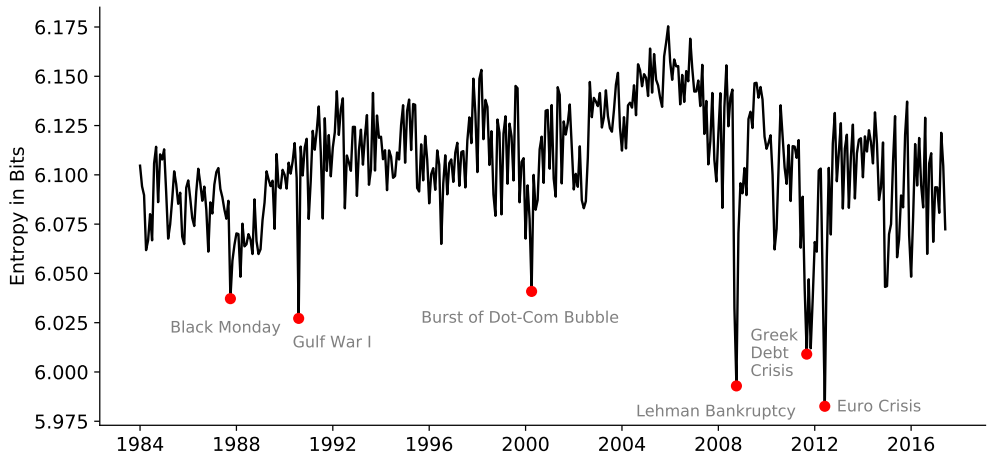
	News Entropy	NE (econ.)	NE (poli.)	NE (cult.)	News Pressure	EPU
News Entropy	1.00					
NE (econ.)	0.55	1.00				
NE (poli.)	0.87	0.13	1.00			
NE (cult.)	0.47	0.33	0.28	1.00		
News Pressure	-0.38	0.02	-0.52	0.05	1.00	
EPU	-0.55	-0.32	-0.50	-0.19	0.39	1.00

Notes: This table presents correlation coefficients between each of the four news entropy series, News Pressure, and the Economic Policy Uncertainty (EPU) index.

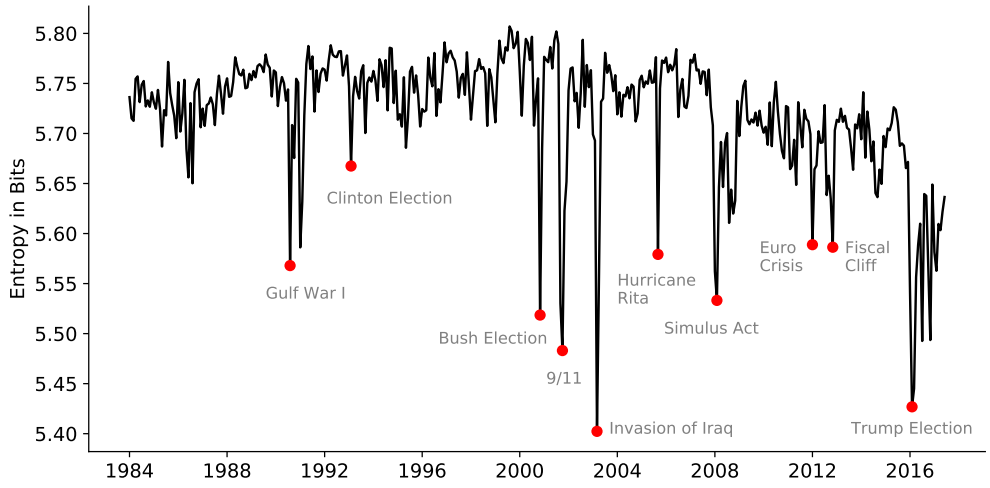
Iraq War, political events such as the US government shutdown in 2013, natural disasters, after the 9/11 terrorist attacks, and close to presidential elections. Strikingly, there seem to be both increased volatility levels and a general downward trend in news entropy starting from 2000 with the 2016 presidential elections representing the overall minimum of the time series.

Next, we construct news entropy measures for thematically related topic subsets. Specifically, based on the topic taxonomy provided by Bybee et al. (2019), we select topics falling into the three broadest categories: economics, politics, and culture. They consist of 77, 59 and 44 topics, respectively. We separately renormalise the topic probabilities for each category and then compute the individual news entropy series. Figure 1.3 shows the respective graphs. We see that the entropy of these subsets picks up the different events seen in the overall graph. More specifically, the news entropy for the economics subset spikes during events such as Black Monday and the Lehman Bankruptcy. Interestingly, the burst of the dot-com bubble is not visibly picked up by the overall news entropy while the economic news entropy series shows a clear negative spike. As expected, political news entropy spikes during events such as presidential elections. Lastly, the culture news entropy series is very noisy and does not seem to pick up any significant events. We additionally compute an inverse Herfindahl index for the topic shares as an additional measure of news diversity. A comparison of the two series can be found in Appendix 1.A, but the main difference between the two is the tails of their respective distributions, with the inverse Herfindahl index featuring a larger degree of negative skewness and positive kurtosis by its nature.

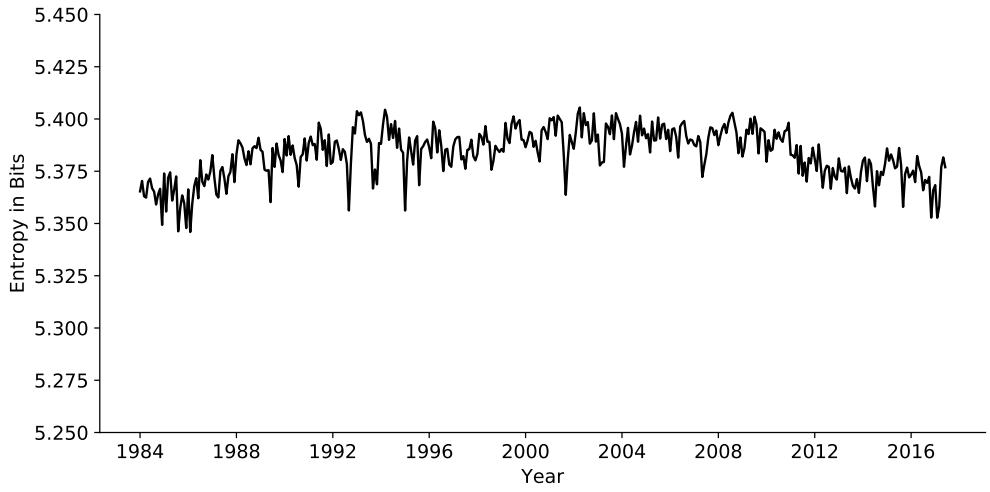
While it may be assumed that the four entropy series are all highly correlated with each other as they might predominantly capture common factors, Table 1.1 shows this is not the case. The only two series which are highly correlated are the main entropy series and the political entropy series, highlighting the dominance of



(a) Economics News



(b) Politics News



(c) Culture News

Figure 1.3. News Entropy for Topic Subsets.

Notes: The figure shows the news entropy series for different topic subsets.

the political news cycle. All other series are positively correlated, but have substantially lower correlation coefficients and thus represent distinct information.

An interesting immediate result is the downward trend in the series which begins following the onset of the financial crisis. Using a Quandt likelihood ratio test (Andrews (2003)) to detect an unknown structural break date, we find one in August 2008. Repeating this procedure for the political entropy measure also results in a structural break being detected in 2008, although the break occurs earlier in the January of that year. We do not find similar evidence of a structural break in the 21st century for the economics or culture series. This result could be interrogated much further, but suggests that news coverage has become more concentrated on a smaller set of dominant topics since the Great Recession, especially with respect to political discourse. A worthwhile point to make is that 2008 was the first year that the Wall Street Journal was under the ownership of News Corp, and it is possible that this shift partially explains the structural break in the newspaper's topic selection.

1.3.3 News Entropy, News Pressure and Policy Uncertainty

News Pressure

We compare news entropy to the concept of news pressure by Eisensee and Strömberg (2007). News pressure measures the amount of airtime a news broadcast allocates to the top three news segments in a day. Specifically, it is defined as the median number of minutes devoted to the first three news segments across broadcasts in a day. The underlying intuition is that the top three news segments represent the most newsworthy events on a given day. Thus, on days of high news pressure – that is, longer airtime for the top three news stories – there is a large amount of newsworthy material and important events dominate the news. Eisensee and Strömberg (2007) show that in turn secondary news get crowded out and receive less coverage. Furthermore, recent applications of news pressure in the field of political economy have shown, for example, that higher news pressure correlates with the likelihood of military attacks (Durante and Zhuravskaya, 2018) and US presidential executive orders (Djourelova and Durante, 2020).

One drawback of news pressure is that it heavily relies on the structure of news broadcasting. In addition, there is no corresponding measure for text-based news reporting. In this context, news entropy provides an alternative measure for newsworthiness based on unstructured news data. Table 1.1 shows the correlations between news pressure and our four news entropy series. We find that news pressure is most strongly correlated with the political news entropy series with a correlation of -0.52. As expected, it is therefore moderately correlated with the overall news entropy series. Interestingly, there is no significant correlation between news pressure and economic news entropy. This result is rather intuitive as it suggests that

economic news are rarely part of the top three news segments on TV. Moreover, we observe that there is no significant correlation between news pressure and the culture news entropy series. Hence, this implies that top news stories on TV are dominated by political events. In future applications, our method could be applied to directly compute the entropy of TV news transcripts.

Policy Uncertainty

Next, we investigate the relationship between news entropy and the Economic Policy Uncertainty (EPU) Index by Baker, Bloom, and Davis (2016). The EPU is an index constructed based on the frequency of the words "uncertain" or "uncertainty" and "economic" or "economy" in newspaper articles in combination with six other policy relevant terms. Similar to our results, Baker, Bloom, and Davis (2016) find that the EPU index spikes near major policy-relevant events. Further empirical applications of the EPU include, for example, Gulen and Ion (2016) who provide evidence of a strong negative relationship between firm-level capital investment and the aggregate level of uncertainty.

Figure 1.A.4 displays the two time series. As documented in Table 1.1, we find that the EPU index and news entropy move in opposite directions with a correlation of approximately -0.55. This suggests that economic uncertainty increases as news entropy decreases. The same holds for the economic and political news entropy series. Interestingly, the correlations of news entropy and political news entropy do not differ much, implying that mostly politics news are associated with policy uncertainty. The correlation between the cultural news entropy series and the EPU is significantly weaker. It is worth noting that there is a positive correlation between the EPU index and news pressure suggesting that a higher frequency of newspapers mentioning uncertainty is associated with longer coverage of the top three stories in TV news broadcasting.

Impulse Responses

To further investigate the relationship between news entropy and commonly used economic uncertainty measures, we rely on the local projection method of Jordà (2005) and include four series as dependent variables: the EPU, the S&P 100 Volatility Index (VXO), and the macroeconomic uncertainty series from Jurado, Ludvigson, and Ng (2015). We directly estimate the impulse response functions via a local projection method which will be explained in more detail subsequently. In our specification, we allow the uncertainty measures to respond contemporaneously to the shock so as not to defeat the object of the exercise. The contemporaneous value of industrial production as well as lags are included, meaning that the shock is identified as a change in entropy that is orthogonal to output. Figure 1.4 shows the results.

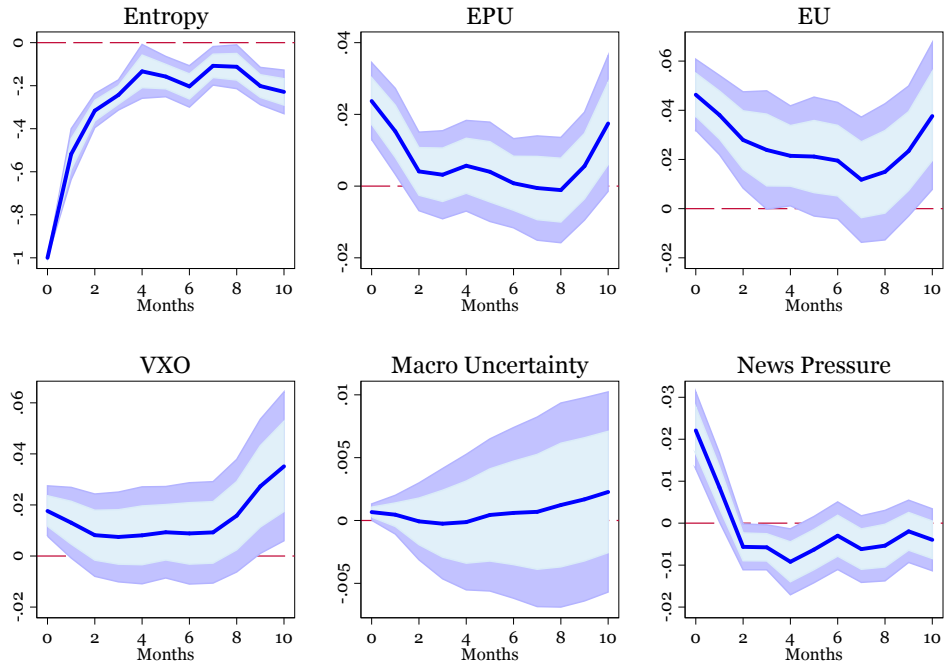


Figure 1.4. Responses of Other Measures to a Fall in Entropy.

Note: The figure shows the impulse response functions of the different measures to a news entropy shock. The light (dark) blue shaded area represents the 68% (90%) Newey-West adjusted confidence intervals.

The VXO, economic and economic policy uncertainty indexes and the news pressure index all display a rise upon impact of the shock and we can reject the null hypothesis of zero impact coefficients for these series at the one percent level. Thus, there seems to be a clear link between entropy and economic uncertainty as we argued previously. The macroeconomic uncertainty index does not respond to any significant degree. Just like the dynamics of entropy, the majority of the series return back to steady state very quickly.

1.4 Economic and Financial Impacts

This section investigates the relationship between news entropy and the economy at the firm, macroeconomic, and financial levels. Section 1.4.1 analyses the impact of changes in news entropy at the firm-level. Section 1.4.2 examines the effects of news entropy shocks on important macroeconomic indicators. Section 1.4.3 examines the relevance of news entropy to asset pricing.

1.4.1 Firm-Level Impact

We examine the firm-level impact of news entropy using option-implied stock price volatility as a proxy for firm-level uncertainty. The data sample contains 136,578 observations on 5,460 firms from 1996 to 2012 obtained from Baker, Bloom, and Davis (2016). Table 1.2 shows the results of quarterly 30-day implied stock price volatility regressed on quarterly average news entropy using firm sales as weights. Columns (1) to (5) rely on the same baseline identification strategy as Baker, Bloom, and Davis (2016) and adopt their measure of firm exposure to uncertainty about government purchases of goods and services.

The specification in column (1) regresses the log of 30-day implied volatility on the logarithm of news entropy. Additionally, the ratio of federal government purchases to GDP is included as a policy control. The coefficient of logged news entropy is highly statistically significant. In this specification, a one percent decrease in news entropy connected to a 21.59% increase in implied volatility. We find that an increase in the ratio of federal purchases to GDP is associated to lower volatility. Column (2) shows the results obtained by Baker, Bloom, and Davis (2016) using the logarithm of the EPU index. Column (3) includes firm and time fixed effects. Additionally, this specification interacts news entropy with firm-level exposure to government purchases. This specification yields a strong relationship between news entropy and implied volatility for firms with greater exposure to government purchases. Column (4) includes the Chicago Board Options Exchange Volatility Index (VIX) in the regression specification. This results in a sign reversal for the news entropy coefficient and a highly significant VIX coefficient. As noted by Baker, Bloom, and Davis (2016) in case of the EPU, this is expected as the VIX measures the 30-day implied volatility on the S&P500 index and should thus be strongly related to the average 30-day implied volatility for publicly listed U.S. firms. Column (5) includes firm and time fixed effects and interacts all regressors with firm-level exposure to government purchase. We find that intensity adjusted news entropy has highly statistically significant coefficient that is larger in magnitude compared to the baseline specification in column (1). We observe that the coefficient on the VIX is statistically indistinguishable from zero. This allows us to draw the same conclusion as the one by Baker, Bloom, and Davis (2016) with respect to the EPU: the VIX has the largest explanatory power for the average firm's 30-day implied volatility. Once we account for exposure to government purchases, however, news entropy explains a significant part of firm-level implied volatility. In summary, the results from running the baseline specifications using news entropy as a predictor for option implied stock price volatility in columns (1) to (5) mirror the findings from Baker, Bloom, and Davis (2016).

Table 1.2. Option-Implied Stock Price Volatility and News Entropy.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log(Entropy)	-21.585***			2.651**					
	(0.995)			(1.04)					
Log(Entropy) × Intensity			-27.098***		-25.549***	-22.022***			
			(7.568)		(8.167)	(7.335)			
Log(VIX)				0.715***					
				(0.013)					
Log(VIX) × Intensity					0.044	0.007			
					(0.09)	(0.115)			
$\frac{\text{Federal Purchases}}{\text{GDP}}$	-14.214***	-19.3***		-8.139***					-14.227***
	(1.531)	(1.5)		(1.481)					(1.531)
$\frac{\text{Federal Purchases}}{\text{GDP}}$ × Intensity			-30.9**		-30.629**	-31.397***	-29.416***	-30.081***	
			(12.421)		(12.309)	(12.131)	(12.389)	(12.335)	
Log(EPU)		0.432***							
		(0.010)							
Log(EPU) × Intensity						0.074		0.094	
						(0.089)		(0.066)	
Log(Entropy Economics) × Intensity							-30.371***	-27.569***	
							(5.55)	(5.343)	
Log(Entropy Politics) × Intensity							3.031	4.986***	
							(1.882)	(1.666)	
Log(Entropy Culture) × Intensity							-17.077	-18.215	
							(21.375)	(21.216)	
Standardised Entropy									-0.08***
									(0.004)
Firm and Time Effects	No	No	Yes	No	Yes	Yes	Yes	Yes	No

Notes: Dependent variable: natural log of the 30-day implied volatility for the firm, averaged over all days in the quarter. The sample is taken from Baker, Bloom, and Davis (2016) and contains 136,578 observations on 5,460 firms from 1996 to 2012. Intensity is a firm's exposure to federal purchases of goods and services. All regressions are weighted by a firm's average sales. Standard errors are clustered by firm.

In addition, we confirm the above findings using a second set of regressions. Column (6) runs the same specification as column (5) with firm and time fixed effects as well as exposure to government purchase but additionally includes the EPU. The news entropy coefficient is significant and of similar magnitude as in the previous specifications. Strikingly, both the coefficient of the EPU and the VIX are statistically indistinguishable from zero while the news entropy coefficient is highly significant. This observation indicates that when comparing the three measures to each other in a setting where we take into account government exposure, news entropy subsumes the effects of the other two, which is in line with its construction as a broader measure. Column (7) simultaneously includes the entropy of the three news subcategories economics, politics and culture in combination with fixed effects and government exposure in place of the general news entropy. We observe that only the economics entropy series has a statistically significant relationship with stock price volatility. Column (8) includes the EPU as a control. The resulting coefficient is not significant. Interestingly, the coefficient of the politics news entropy series is now statistically significant with a positive coefficient. That is, once we control for the use of the words such as "uncertainty" as measured by the EPU, stock price volatility increases in political news entropy. Finally, we note that when using the logarithm of news entropy, a one percent change in news entropy is rather large looking at the entire time series. This is in contrast to the EPU as the EPU is a normalised unitless index while entropy is measured in bits. To test whether this affects our results, we provide an alternative specification where we measure the effect of a one-standard-deviation change in news entropy. While we find that the resulting coefficient is much smaller in magnitude as expected, it is highly statistically significant. At the same time, the coefficient of the ratio of federal government purchases to GDP is virtually unchanged. Hence, this indicates that our results do not depend on the normalisation method and confirms the above findings.

1.4.2 Macroeconomic Impact

We now investigate the relationship between important economic indicators and news entropy. As a preliminary exercise, we look at the cyclical properties of the set of measures by examining the correlation of each with industrial production (IP). After each series is detrended with a Hodrick-Prescott filter (with a smoothing parameter of 129,600), the correlation between the main entropy series and the log of IP is a mere 0.03 and is not statistically significant at conventional levels. Interestingly, the series is therefore acyclical. This is a notable difference from the EPU, which exhibits strong and significant countercyclicality, with a correlation coefficient of -0.35 with the log of IP. The topic entropy measures also all display a lack of any

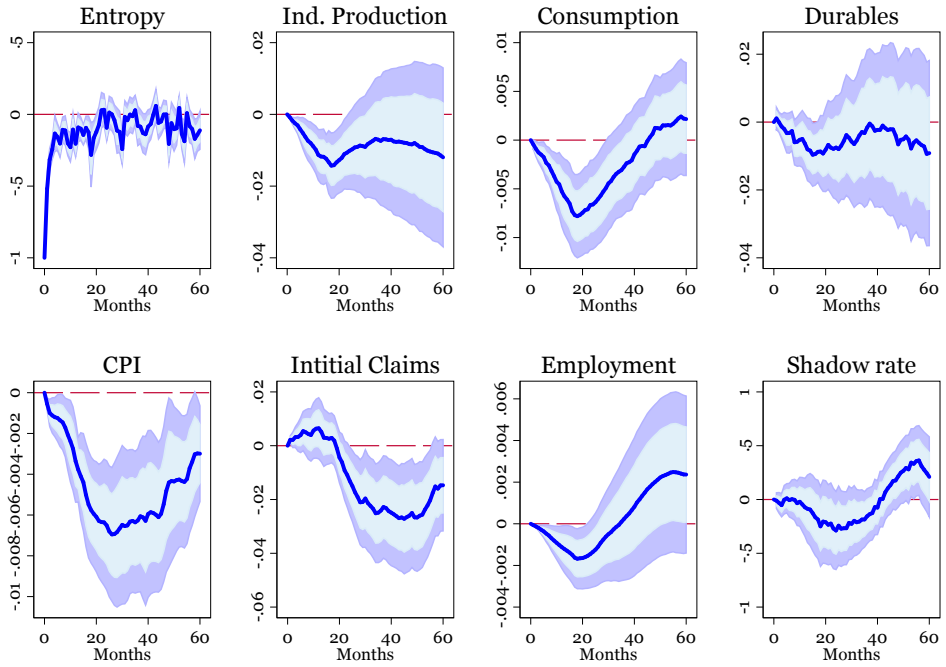


Figure 1.5. Responses to News Entropy Shock.

Note: The figure displays the estimated impulse response functions of the endogenous variables for a shock to the entropy measure. The light (dark) blue shaded area represents the 68% (90%) Newey-West adjusted confidence intervals.

kind of cyclical pattern – the politics measure is the only series whose correlation coefficient is statistically significant.

Next, we use the local projection method of Jordà (2005) to fully explore the macroeconomic impacts of a shock to the entropy measure, directly estimating the impulse response functions (IRFs). See Plagborg-Møller and Wolf (2019) for a full review of this approach as well as its similarities and differences with the structural vector autoregression (SVAR) approach. The specification for the local projection can be expressed as

$$Y_{t+h} = \alpha_h + \gamma_h e_t + \psi_h(L)Z_t + u_{t+h}$$

where Y is an endogenous variable of interest, e_t is the main entropy measure in period t and Z_t is a set of control variables. The endogenous variables we investigate include industrial production, non-durable consumption and services, durable consumption, initial claims for unemployment insurance, hours worked, the consumer price index (CPI) and the shadow Federal Funds rate from Wu and Xia (2016). All variables except the last enter in log levels. The set of controls in each regression

includes six lags of the entropy measure, the current value and six lags of the dependent variable and the current value and six lags of industrial production. A linear trend is also included. As shown by Plagborg-Møller and Wolf (2019), this procedure is equivalent to ordering the entropy measure last in a recursively ordered SVAR and can be considered conservative as such. We stress that we do not endow the news entropy shock with a structural interpretation, as the measure will clearly be a function of a number of underlying shocks. Instead, we interpret the IRFs as capturing, on average, how the set of variables respond after a change in news entropy.

The entropy measure is standardised to have a mean of zero and a standard deviation of one and we examine a negative shock, that is, a fall in entropy. The maximum value of h is set at 60 for a five-year horizon for the IRFs. To correct for serial correlation in the errors, Newey-West standard errors are employed with automatic bandwidth selection (Newey and West, 1994). The sample period is from January 1984 to June 2017.

Figure 1.5 displays the estimated impulse response functions for the endogenous variables along with one standard error confidence bands. Entropy decreases but bounces back almost immediately and does not persistently stay below trend. The shock is contractionary, with output remaining persistently below steady state afterwards. The recession is particularly concentrated in non-durable consumption and services, which exhibits a v-shaped decline, and is also accompanied by a clear decline in the price level. The decline in both of these variables is precisely estimated. Durable consumption falls although the estimates are imprecise. Initial claims increase, indicating a rise in layoffs, while employment falls. This decrease in employment is followed by an overshoot after around three years, mirroring the same pattern found in Bloom (2009) after an uncertainty shock. The shadow Fed. Funds rate falls slightly, which suggests that the Federal Reserve responds according to its Taylor rule in an attempt to counteract the impact by cutting interest rates.

To ensure the contractionary effect of a decrease in entropy is a robust result, we also estimate impulse response functions from an array of modified specifications which include:

- Hodrick-Prescott filtered variables with a smoothing parameter of 129,600.
- three lags of all variables.
- twelve lags of all variables.
- lags of stock prices as an additional control variable.
- lags of the VXO as an additional control variable.
- the contemporaneous value and lags of employment as an additional control variable.

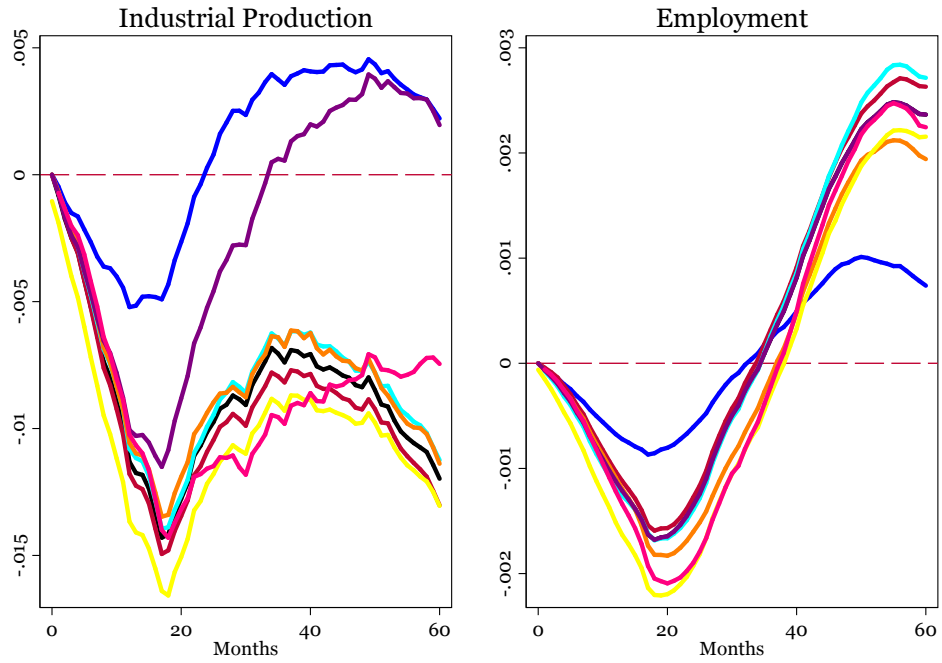


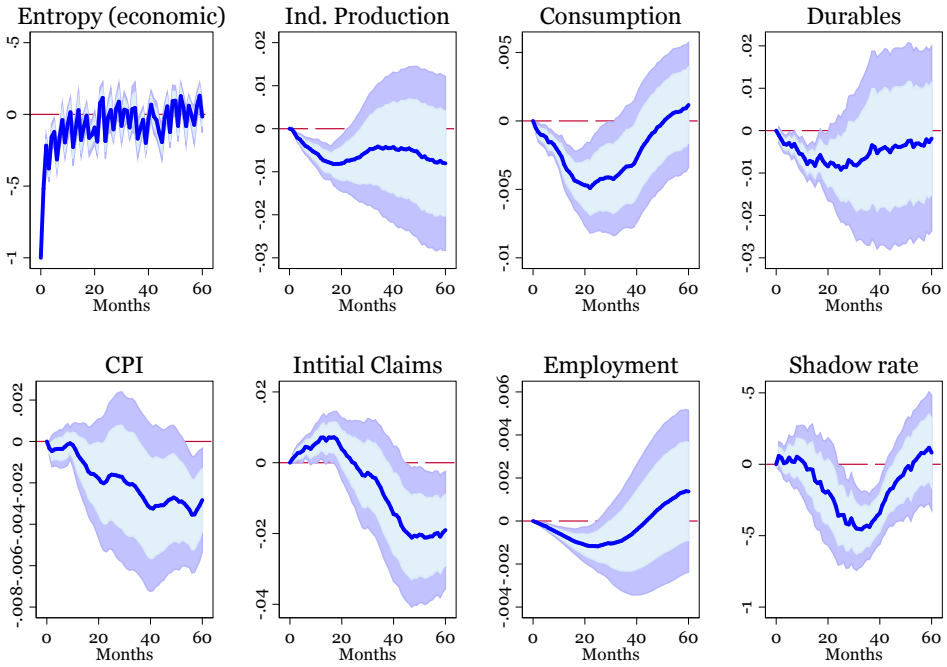
Figure 1.6. Robustness Tests for the News Entropy Shock.

Note: The figure shows the estimated impulse response functions from alternative specifications represented by different colours in addition to the baseline specification (black) which include: Hodrick-Prescott filtered variables (blue), three lags of all variables (dark red), twelve lags of all variables (cyan), stock prices as a control (orange), lags of the VXO as an additional control variable (pink), employment as a control (purple). The yellow line displays the estimates from the specification with an alternative causal ordering.

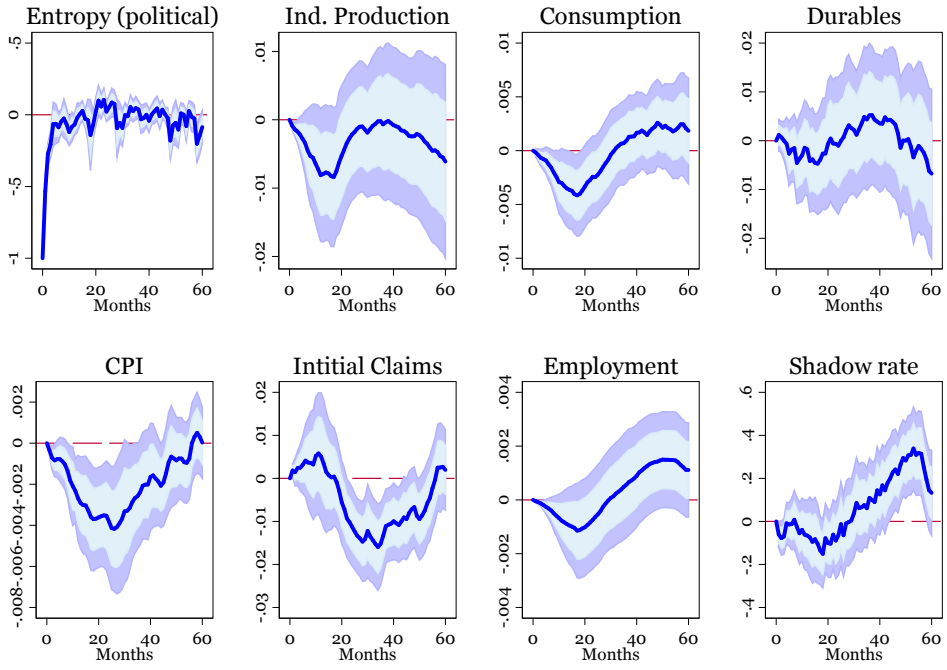
- using a different causal ordering equivalent to ordering the entropy measure first in a recursively identified SVAR.

The estimated impulse response functions for each of these alternative specifications is presented in Figure 1.6. The contractionary response remains present in all specifications, and most of them yield extremely similar estimates to the baseline specification. Using a Horick-Prescott filter results in industrial production displaying the overshoot pattern exhibited by employment, while the addition of employment as a control variable slightly attenuates the response after the 18 month horizon.

We next investigate shocks to the topic-specific measures of entropy. The estimated IRFs can be found in Figure 1.7. The main finding from these is that shocks to the political and economic entropy series are also contractionary, and lead to a qualitatively similar (but quantitatively smaller) decline in industrial production. The response of consumption and the monetary policy variable is particularly pronounced for the economic series.



(a) Responses to Economic News Entropy Shock.



(b) Responses to Political News Entropy Shock.

Figure 1.7. Responses to Entropy Shock of News Themes.

Notes: The figures show the impulse response functions for economic and political news entropy shocks. The light blue shaded area represents the 68% Newey-West adjusted confidence intervals. The dark blue shaded area represents the 90% Newey-West adjusted confidence intervals.

We also look at large changes in entropy by defining an indicator variable that takes the value of 1 when the entropy measures is more than one standard deviation below its mean. 58 such months in the sample are classified as low entropy periods. We then include this indicator in the local projection, keeping the rest of the specification the same. The estimated IRFs for this shock are shown by Figure 1.A.2 in Section 1.A. They closely resemble the benchmark IRFs, with a contraction occurring as well as notable deflation. Additionally, we consider the inverse Herfindahl index as a measure of news diversity in the local projections. Figure 1.A.3 displays the IRFs in this case, and once again they look very similar to those in Figure 1.5.

Nonlinear Effects

Next, we further our analysis by exploring whether there are nonlinearities present in the impulse responses of the macroeconomic variables to an entropy shock. We have previously noted that many of the large decreases in the news entropy measure corresponded to natural or economic disasters such as Black Monday, 9/11 and the collapse of Lehman Brothers. A natural question that arises is then whether larger news entropy shocks have a disproportionate impact on the macroeconomy. To investigate this, we run the following specification of the local projection with the same set of dependent variables as previously:

$$Y_{t+h} = \alpha_h + \gamma_h e_t + \bar{\gamma}_h e_t^2 + \tilde{\gamma}_h e_t^3 + \psi_h(L) Z_t + u_{t+h}$$

We therefore allow for nonlinearities in the impulse response function via the inclusion of the quadratic and cubic terms in the shock. We compare these IRFs estimated from the nonlinear LP to those estimated from the benchmark linear specification in Figure 1.8. This clearly illustrates the presence of nonlinearities, as substantial deviations between the two IRFs are present. Crucially, the response of entropy in the two specifications is very similar. A key difference in IRFs the nonlinear specification is the tendency of most variables to display a sharper contractionary movement than in the linear case, but then a rebound that involves a sizeable expansion after around two years. For example, in the linear specification, the estimate IRF for layoffs (as measured by initial claims for unemployment insurance) is more or less flat over the horizon period, whereas in the nonlinear specification the variable displays a large rise in response to the shock, which is then followed by a substantial fall. A similar phenomenon is present in Bloom (2009), who also estimates this rebound for many variables after an uncertainty shock. Another key difference is that the response of monetary policy is found to be much more pronounced in the nonlinear specification. This may suggest that the Federal Reserve is taking more drastic action in response to these disasters, as was the case with Quantitative Easing (QE) during the financial crisis.

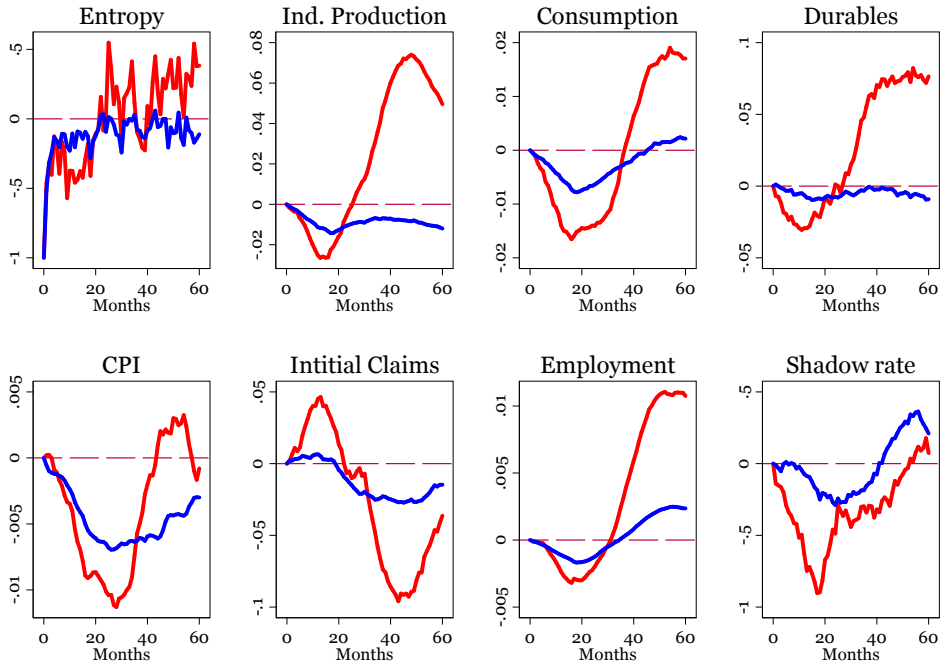


Figure 1.8. Impulse response functions from nonlinear local projections

Note: The red line in each figure corresponds to the impulse response function estimated from the specification which includes higher order terms of the shock. The blue line in each figure corresponds to the benchmark specification which is linear in the shock.

1.4.3 Financial Impact

Next, we investigate the relevance of our news entropy variable in an asset pricing context. Specifically, we pose the question: is news entropy priced in the cross-section of returns? To do this, we implement the canonical method of Fama and MacBeth (1973) to estimate linear factor models. Let J denote the total number of portfolios and T denote the total number of time periods used in the estimation. The procedure involves first running J time-series regressions of the form

$$R_t^{e,j} = a_j + \beta_j f_t + \epsilon_t^j \qquad j = 1,...,J$$

where $R_t^{e,j}$ is the excess return (over the risk-free rate) of asset j in period t and f_t is a $K \times 1$ vector of factors. The second step of the procedure estimates the risk price for each factor by using the first-stage estimated factor loadings and running T cross-sectional regressions

$$R_t^{e,j} = \lambda_t \hat{\beta}_j + \alpha_t^j \qquad t = 1,...,T$$

Table 1.3. Fama and MacBeth (1973) Regressions.

λ_{NE}	$\lambda_{NE(econ.)}$	$\lambda_{NE(poli.)}$	λ_{R^M}	λ_{SMB}	λ_{HML}	MAPE
0.079						1.39
(4.32)						
	0.034					1.42
	(4.26)					
		0.301				2.79
		(4.37)				
			0.917	0.028	0.326	0.99
			(4.13)	(0.18)	(2.17)	

Notes: The table reports results of Fama and MacBeth (1973) regressions for the 25 Fama-French portfolios. See text for full estimation details. MAPE denotes the mean absolute pricing error. Square brackets denote t-statistics. The sample period is from January 1984 to June 2017.

The estimated risk factor prices are then given by

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\lambda_t.$$

As test assets, we follow the majority of the literature and use the 25 Fama and French (1993) portfolios sorted by size and book-to-market. We estimate three single-factor models with the main news entropy measure, the economic news entropy measure and the political news entropy measure. As a benchmark with which to draw a comparison, we also estimate the Fama and French (1993) three factor model with the excess return on the market portfolio, the size premium (SMB) and the value premium (HML). We report the estimated risk prices from each model as well as the t-statistics. Additionally, we also report the mean absolute pricing error from each model, which indicates how effectively each model can explain the overall cross-section of returns.

Table 1.3 displays the results from the Fama-MacBeth regressions. The first notable result is that all three entropy measures display positive risk prices which are statistically significant at conventional levels. Assets which are more exposed to the news cycle earn a risk premium. For the main news entropy measure, a one standard deviation in exposure (β) is associated with a 3.04 percentage point increase in the annualised expected excess return on an asset. This value is very similar for the economic news entropy measures at 2.44 percentage points. Interestingly, for the political news entropy it is more than double at 5.36 percentage points. The pricing errors are lowest for the main news entropy single-factor model, although they remain low across all three models. Comparison to the three-factor Fama-French model estimates reveals that the entropy models are able to achieve a comparable level of performance, with only slightly higher pricing errors.

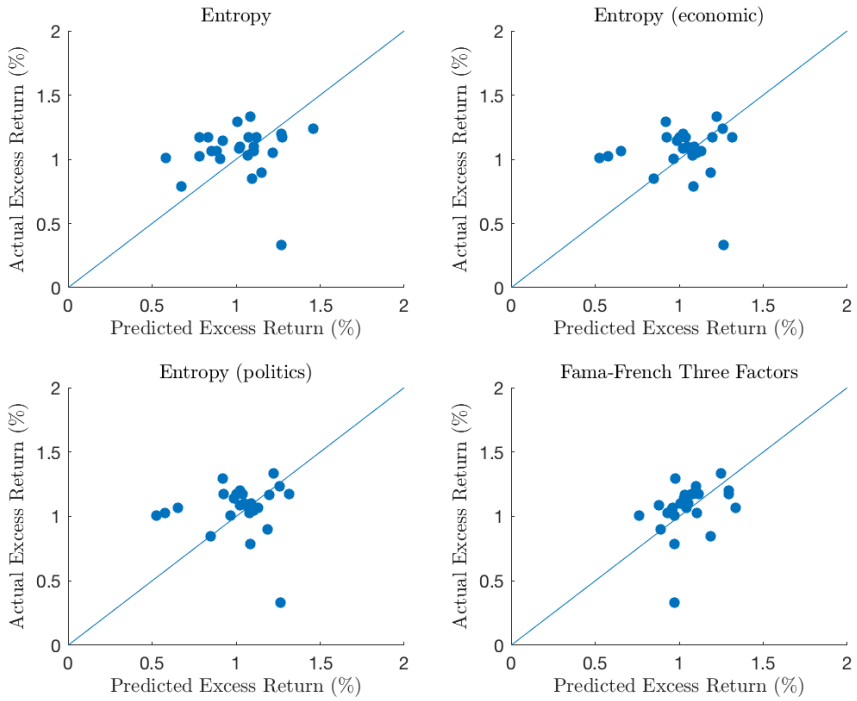


Figure 1.9. Fama-MacBeth Plots

Note: The figure plots the predicted excess return for each of the 25 Fama-French portfolios against its sample average excess return for each of the four factor models.

Figure 1.9 plots the predicted excess return on each portfolio from each of the four factor models against the actual expected excess return. This further illustrates the success of the news entropy factor models, as the pricing error for most portfolios is low. All four models struggle to successfully price the small growth portfolio in this sample period.

1.5 Conclusion

We have introduced the concept of news entropy to parsimoniously characterise the complex structure of news content in simple terms. Empirically, we find that news entropy features negative skewness and positive kurtosis, as it collapses during times of significant political and economic unrest as well as natural disasters such as hurricane Rita. We find that these decreases in news entropy coincide with periods of high uncertainty, and results from local projections demonstrate that they are followed by a macroeconomic contraction. Meta-topic specific analysis shows that economic news entropy has a particularly strong association with these dynamics, with the relationship less strong for political news entropy. Allowing for nonlinearities in the

impulse response functions substantially alters their shape, resulting in a deeper contraction but then a strong rebound and overshoot, dovetailing with the discussion of a V-shaped recession during the early parts of the SARS-CoV2 pandemic. While we do not yet have the required data to update our news entropy measure through to the ongoing SARS-CoV2 pandemic of 2020, this represents a time of unprecedented levels of both uncertainty and news concentration, with global news coverage focused almost entirely on one topic. This crisis thus acts a clear illustration of our central concept. In future work we plan to document the evolution of news entropy during the pandemic, and to examine the macroeconomic and financial ramifications this had.

Our measure currently only exists for the United States, but a key benefit of our method is how effectively it generalises to news media in other countries, potentially written in other languages. We therefore intend to create news entropy measures for a range of countries, which would allow us to assess whether the impact of news entropy varies internationally.

Lastly, from a methodological perspective, we provide a new, flexible framework that builds on the combination of probabilistic machine learning techniques and information-theoretic concepts. This approach can be adapted to a variety of other probabilistic models to construct economic measures from unstructured data sources in a theoretically well-defined manner.

Appendix 1.A Additional Figures

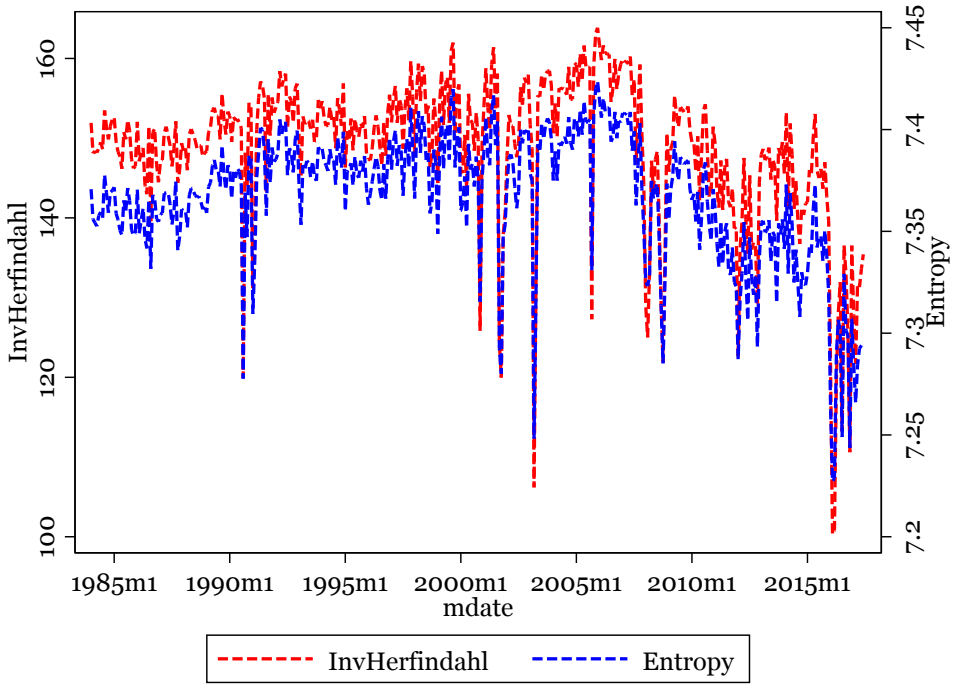


Figure 1.A.1. Comparison of News Entropy and Inverse Herfindahl Index.

Note: The blue dashed line represents news entropy, while the red dashed line represents the inverse Herfindahl index.

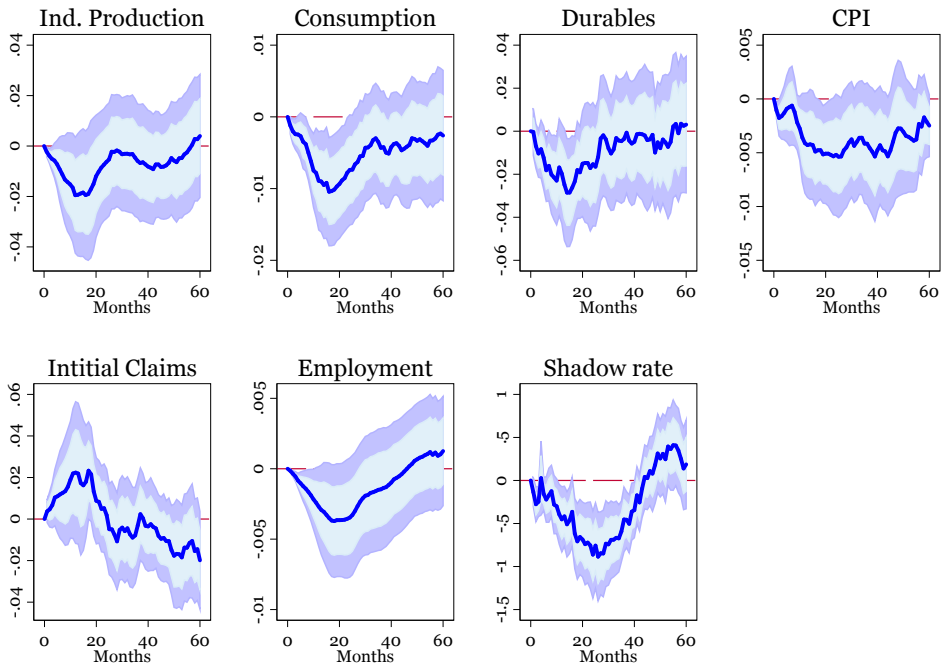


Figure 1.A.2. Responses to News Entropy Indicator Shock.

Note: The light blue shaded area represents the 68% Newey-West adjusted confidence intervals. The dark blue shaded area represents the 90% Newey-West adjusted confidence intervals.

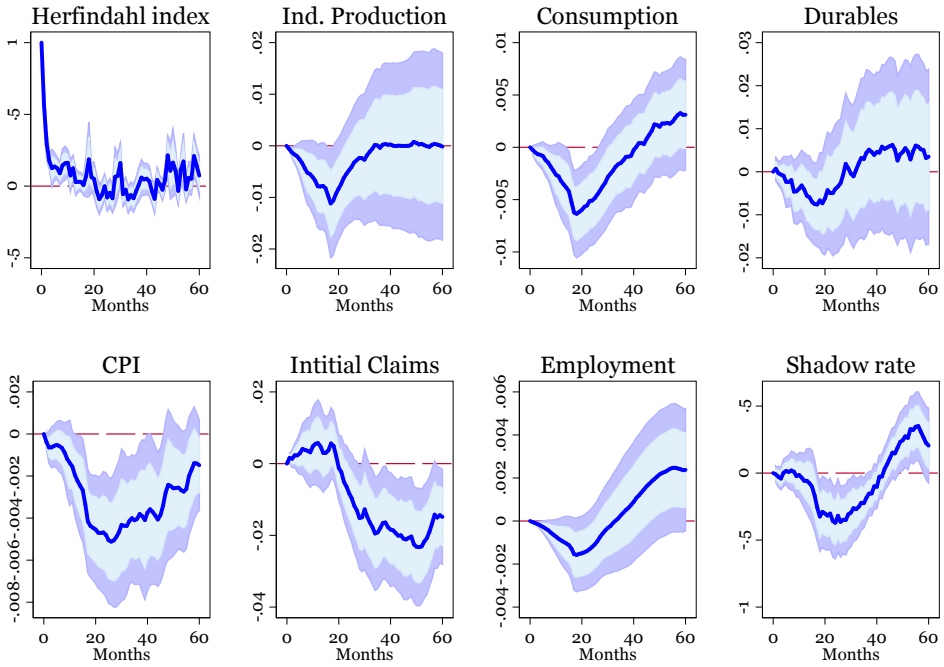


Figure 1.A.3. Responses to Inverse Herfindahl Index Shock.

Note: The light blue shaded area represents the 68% Newey-West adjusted confidence intervals. The dark blue shaded area represents the 90% Newey-West adjusted confidence intervals.

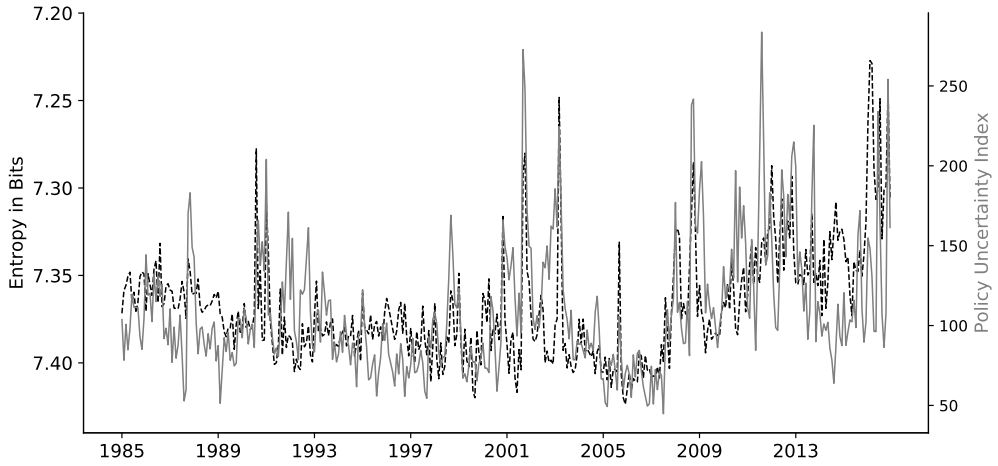


Figure 1.A.4. News Entropy and Policy Uncertainty.

Note: This figure shows our measure and the Policy Uncertainty Index from 1985 to 2016. The ordinate for the news entropy has been inverted to allow for an easier visual comparison.

References

Andrews, Donald W. K. 2003. "Tests for Parameter Instability and Structural Change with Unknown Change Point: A Corrigendum." *Econometrica* 71 (1): 395–97. URL: https://ideas.repec.org/a/ecm/emetrp/v71y2003i1p395-397.html. [15]

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." *Quarterly Journal of Economics* 131 (4): 0–52. DOI: 10.1093/qje/qjw024.Advance. arXiv: 9809069v1 [arXiv:gr-qc]. [6, 8, 12, 16, 18, 19]

Barro, Robert J. 2006. "Rare Disasters and Asset Markets in the Twentieth Century." *Quarterly Journal of Economics* 121 (3): 823–66. URL: https://ideas.repec.org/a/oup/qjecon/v121y2006i3p823-866..html. [7]

Bertsch, Christoph, Isaiah Hull, and Xin Zhang. 2021. "Narrative fragmentation and the business cycle." *Economics Letters* 201: 109783. DOI: 10.1016/j.econlet.2021.109783. [7]

Blei, David M., and John D. Lafferty. 2007. "A correlated topic model of Science." *Annals of Applied Statistics* 1 (1): 17–35. DOI: 10.1214/07-AOAS114. arXiv: 0712.1486. [11]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1. [6–8]

Bloom, Nicholas. 2009. "The Impact of Uncertainty Shocks." *Econometrica* 77 (3): 623–85. DOI: 10.3982/ecta6248. [22, 25]

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. "The Structure of Economic News." *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3446225. [7, 12, 13]

Chahrour, Ryan A., Kristoffer Nimark, and Stefan Pitschner. 2019. "Sectoral Media Focus and Aggregate Fluctuations." *SSRN Electronic Journal*, 1–42. DOI: 10.2139/ssrn.3477432. [8]

DeDeo, Simon, Robert X.D. Hawkins, Sara Klingenstein, and Tim Hitchcock. 2013. "Bootstrap methods for the empirical study of decision-making and information flows in social systems." *Entropy* 15 (6): 2246–76. DOI: 10.3390/e15062246. [11]

Djourelova, Milena, and Ruben Durante. 2020. "Media Attention and Strategic Timing in Politics: Evidence from U.S. Presidential Executive Orders." [15]

Durante, Ruben, and Ekaterina Zhuravskaya. 2018. "Attack When the World Is Not Watching? US News and the Israeli-Palestinian Conflict." *Journal of Political Economy* 126 (3): DOI: 10.1086/697202. [15]

Eisensee, Thomas, and David Strömberg. 2007. "News droughts, news floods, and U. S. disaster relief." *Quarterly Journal of Economics* 122 (2): 693–728. DOI: 10.1162/qjec.122.2.693. [6, 8, 15]

Fama, Eugene F., and Kenneth R. French. 1993. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1): 3–56. DOI: 10.1016/0304-405X(93)90023-5. arXiv: arXiv:1011.1669v3. [27]

Fama, Eugene F., and James MacBeth. 1973. "Risk , Return, and Equilibrium : Empirical Tests." *Journal of Political Economy* 81 (3): 607–36. [26, 27]

Fligstein, Neil, John Stuart Brundage, and Michael Schultz. 2014. "Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of "Macroeconomics" as a Sense making and Cultural Frame." *IRLE Working Paper* 111 (14): [7]

Gabaix, Xavier. 2012. "Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance." *Quarterly Journal of Economics* 127 (2): 645–700. DOI: 10.1093/qje/qjs001. [7]

Gulen, Huseyin, and Mihai Ion. 2016. "Policy uncertainty and corporate investment." *Review of Financial Studies* 29 (3): 523–64. DOI: 10.1093/rfs/hhv050. [16]

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2017. "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach." *Quarterly Journal of Economics* 133 (2): 801–70. DOI: 10.1093/qje/qjx045. [7]

Jordà, Òscar. 2005. "Estimation and inference of impulse responses by local projections." *American Economic Review* 95 (1): 161–82. DOI: 10.1257/0002828053828518. [6, 16, 21]

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. 2015. "Measuring uncertainty." *American Economic Review* 105 (3): 1177–216. DOI: 10.1007/s11225-009-9207-0. [16]

Larsen, Vegard H., and Leif A. Thorsrud. 2019. "The value of news for economic developments." *Journal of Econometrics* 210 (1): 203–18. DOI: 10.1016/j.jeconom.2018.11.013. [7]

Mahajan, Anuj, Lipika Dey, and Sk Mirajul Haque. 2008. "Mining financial news for major events and their impacts on the market." *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008*, 423–26. DOI: 10.1109/WIIAT.2008.309. [7]

Manela, Asaf, and Alan Moreira. 2017. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123 (1): 137–62. DOI: 10.1016/j.jfineco.2016.01.032. [8]

McCracken, Michael W., and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business and Economic Statistics* 34 (4): 574–89. DOI: 10.1080/07350015.2015.1086655. [12]

Murdock, Jaimie. 2019. "Topic Modeling the Reading and Writing Behavior of Information Foragers." (June): arXiv: 1907.00488. URL: http://arxiv.org/abs/1907.00488. [10]

Newey, Whitney K., and Kenneth D. West. 1994. "Automatic lag selection in covariance matrix estimation." *Review of Economic Studies* 61 (4): 631–53. DOI: 10.2307/2297912. [22]

Nimark, Kristoffer P. 2014. "Man-Bites-Dog Business Cycles." *American Economic Review* 104 (8): 2320–67. DOI: http://dx.doi.org/10.1257/aer.104.8.2320. [7]

Nimark, Kristoffer P., and Stefan Pitschner. 2019. "News media and delegated information choice." *Journal of Economic Theory* 181: 160–96. DOI: 10.1016/j.jet.2019.02.001. [7]

Peress, Joel. 2014. "The media and the diffusion of information in financial markets: Evidence from newspaper strikes." *Journal of Finance* 69 (5): 2007–43. DOI: 10.1111/jofi.12179. [8]

Plagborg-Møller, Mikkel, and Christian K Wolf. 2019. "Local Projections and VARs Estimate the Same Impulse Responses." [21, 22]

Rauh, Christopher. 2019. "Measuring Uncertainty at the Regional Level Using Newspaper Text." [7]

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (4): 623–56. DOI: 10.1002/j.1538-7305.1948.tb00917.x. [6, 9]

Shannon, C. E. 1951. "Prediction and Entropy of Printed English." *Bell System Technical Journal* 30 (1): 50–64. DOI: 10.1002/j.1538-7305.1951.tb01366.x. [10]

Wachter, Jessica A. 2013. "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" *Journal of Finance* 68 (3): 987–1035. DOI: 10.1111/jofi.12018. [7]

Wu, Jing Cynthia, and Fan Dora Xia. 2016. "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound." *Journal of Money, Credit and Banking* 48 (2-3): 253–91. DOI: 10.1111/jmcb.12300. [21]

Chapter 2

Exploration and Exploitation in US Technological Change*

Joint with Vasco M. Carvalho and Mirko Draca

2.1 Introduction

Technological change and innovation are central to the process of economic growth but are difficult to measure. Following Griliches (1990), efforts to measure technological change and innovation can be summarised according to whether they involve either information on innovation outputs (for example: patents and scientific papers) and inputs (for example: R&D, employment counts of scientists and engineers) as proxy indicators, or are based on the residual information about factor usage that is represented by total factor productivity (TFP).

These approaches face clear challenges when it comes to capturing qualitative change in the range and conceptual basis of technologies over time, as well as the experimental nature of many technological investments. At a fundamental level, innovation choices involve a trade-off between exploration and exploitation. Specifically, a firm may shift between 'exploiting' a breakthrough by developing a given technology in more depth or dedicating more effort to searching and experimenting in a new technological domain. This latter process of search can be characterised as continuing 'exploration'. The trade-off between exploration and exploitation has

* We thank Tiago Cavalcanti, Petra Geraats, Bill Janeway, Alexei Onatski, Christopher Rauh, Mikhail Safronov, and Weilong Zhang for helpful comments and suggestions. We thank participants at the Barcelona Graduate School of Economics Machine Learning for Economics Summer Forum 2019, Turing Research Showcase 2020 at The Alan Turing Institute, Cambridge Applied Micro Seminar and Monash-Warwick-Zurich Text-as-Data Workshop for useful feedback and suggestions. Kuhlen gratefully acknowledges the financial support of The Alan Turing Institute under research award No. TU/C/000030.

been a prominent feature of behavioural theories of the firm, following, for example, Cyert and March (1963) and March (1991).

In parallel with this, there is also a longstanding literature on firm size and innovation focused on the Schumpeterian 'Mark I versus Mark II' debate about the role of large firms in innovation over time. A central question here has been whether larger firms inherently tend towards producing incremental rather than radical innovations (Cohen, 2010; Nicholas, 2015). Given the economies of scale that are associated with size, a shift towards incremental innovation is compatible with firms entering 'exploitation' phases in their growth.

A further literature has discussed scientific and artistic creativity over the individual life-cycle. Creativity is widely thought to peak between the ages of 30 and 40 across a number of domains (Dennis, 1956; Lehman, 1960; Galenson and Weinberg, 2000; Jones, Reedy, and Weinberg, 2014). Recent work studying this question in the context of US patenting (Kaltenberg, Jaffe, and Lachman, 2021) is in line with this, finding that patenting rates peak around the early 40s and that measures of the quality or importance of patenting decline with age.

In this paper, we follow the exploration versus exploitation perspective on innovation and outline new empirical measures that render tangible how a firm or inventor moves through their 'knowledge space'. We implement this approach across US firms, inventors and counties, which we refer to as 'units' of innovation. Our principal contribution is to construct a new empirical measure of unit-level innovation from the corpus of patent texts. The measure that we put forward is based on the changes in the 'text information' implicit in a unit's patent portfolio. As such, it is distinct from and complements existing measures of innovation that are based on inputs, outputs or TFP.

We use unsupervised learning methods to measure shifts in a unit's patenting activities, defined in terms of topics that correspond to probability-weighted word clusters. In short, we identify phases of exploration by measuring how a unit moves across the 'topic space' of its patents. Bigger jumps across the topic space are identified as phases of heightened exploration while stable years are indicative of phases of exploitation.

More specifically, to measure exploration, we first use Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) as a dimension reduction tool that allows us to describe patent texts in terms of a latent topic structure. Applying LDA to a unit-level patent corpus yields two main elements. These are, firstly, a set of endogenously generated knowledge topics and, secondly, a distribution of these topics over the set of a unit's patents.

We then use a 'Bayesian Surprise' (Itti and Baldi, 2009) measure to quantify the extent to which the patents of a given unit in a particular year contain a new mixture of topics compared to what came before. The Bayesian Surprise concept is grounded

in information theory and results in a measure that is defined according to informational 'bits'. The concept has general applicability across social and natural science settings. For example, Itti and Baldi (2009) show that Bayesian Surprise captures real cognitive processes as it predicts what a subject shifts their attention and gaze towards. In our application, the unit-specific past topic distributions function as a prior, to be compared to the topic distribution in the current period. In this way, we use Bayesian Surprise to evaluate how exploratory a unit is at different points in time according to movements across its latent topic space.

We build on this further to construct a measure of 'successful' exploration by adopting the resonance measure proposed by Barron, Huang, Spang, and DeDeo (2018). In short, this measure hinges on how exploration in the current period is different relative to past and future exploration. A unit might move into a different area of its underlying topic space but may not stay in this area. This would be an example of 'unsuccessful' exploration. In contrast, successful exploration is defined as episodes where exploration in the current period is different to past exploration but similar to subsequent firm innovation activity. 'Successful' exploration is therefore an episode of exploration that 'sticks' and is manifested in a lasting change in a unit's underlying topic distribution.

Our empirical implementation of this approach uses a database built up from a match of US Patents and Trademark Office (USPTO) records on the abstracts of patents to information on firms, inventors and counties. This provides us with a long time period for studying the evolution of these units. For firms, we are able to measure exploration behaviour for the period since 1920 while for counties and inventors we study the periods since 1947 and 1976, respectively.

Findings

We implement our exploration measure at a range of levels with a specific focus on identifying developmental patterns in the progress of exploration. We first demonstrate our methodology with a case study of the International Business Machines (IBM) corporation, a firm that was central to the development of computing technology during the 20th century. This case study shows how IBM underwent a major transition from mechanical and analogue to digital technologies in a period centred on the 1950s. This transition is apparent from the basic word frequencies of IBM's patents across decades, the underlying topic structure of the firm's patent portfolio, and from the summary exploration measures that we calculate.

We next look at patterns of exploration across all available firms. Using a measure of 'cumulative exploration' (in effect, the integral of annual exploration flows) we are able to trace out developmental patterns in a firm's innovation behaviour.

That is, there are clear phases of faster and slower exploration, including evidence of widespread 'S-shaped' diffusion-style trajectories.

Interestingly, the principal explanatory factor for these firm exploration trajectories is firm age. The correlation with firm age actually dominates as an explanatory variable when it is included alongside firm size variables and a patent stock measure. There is also a clear 'wedge' between the exploration-age profile of firms versus the firm size-age profile. Practically, this means that exploration tapers off with age faster than firm size, hinting again at a potential developmental pattern in firm behaviour. This is complemented by a pattern of sharply declining Research and Development (R&D) intensity in firm age. On average, the early years of a firm's life in the US data we examine seem to be dedicated to (relatively) more pronounced exploratory and R&D-intensive innovation.

In the final part of our analysis of firms we examine the association between our exploration measure and firm sales growth. This indicates that there is an association that is robust to industry trends and controls for the growth of patenting. Furthermore, the association also holds when controlling for age, indicating that the intensive margin of exploration across firms of the same age has explanatory power. Our measure of successful exploration also appears to be effective at identifying phases of exploration that are more strongly associated with sales growth than the 'general' measure of exploration.

Our next set of findings focuses specifically on the geographical distribution of ICT patenting and exploration across US counties. We observe that exploration is more geographically concentrated than actual patenting itself, but that there is a limited overlap in the concentration of patenting. That is, exploration is occurring away from the main hubs of patenting, with the top examples of this intensive 'periphery' exploration being counties where defense contracting firms have a strong presence. Overall, we find that the concentration of exploration was highest in the period between 1960-1980. The decline following 1980 then occurs alongside an increase in the concentration of ICT patenting itself, in this case towards classic innovation hubs such as Palo Alto.

The final application we look at relates to inventor age and exploration. As discussed, there is a broad literature that has found support for the idea that creativity and scientific productivity peak at middle age. Our findings are in line with this literature. We find that exploration peaks at around the age of 40 across a number of subsets of inventors – the full sample plus the 'superstars' in the top 1% and 0.1%. There are indications that the superstars defined in terms of the volume of patents produced go through 'waves' of exploration but a conventional, middle-aged peak holds for superstars identified according to average lifetime exploration. The life-cycle peaks in exploration are also substantial: inventors are around twice as exploratory at their peak than they are at other periods of life.

Related Literature

In addition to the work on firm growth, inventor life-cycles and economic geography that we have discussed this paper contributes to the emerging literature on using text-based information to measure innovation. Kelly, Papanikolaou, Seru, and Taddy (2018) construct a measure of 'breakthrough patents' using historical USPTO data and following a principle of 'backward importance'. That is, breakthrough patents are those that are amongst the first to feature n-gram phrases that became more common in later patents. Bussy and Geiecke (2020) follow the same intuition of comparing patent similarity across past and future periods but with an implementation focused on Latent Semantic Analysis (LSA) methods. The identification of new or fast-growing ideas in patents is also at the centre of the contributions by Balsmeier, Assaf, Chesebro, Fierro, Johnson, et al. (2018), Bowen, Fresard, and Hoberg (2021) and Packalen and Bhattacharya (2015). Arts, Cassiman, and Gomez (2018) provide a deep discussion of the measurement of patent text similarity, with the additional element of introducing expert (human) validation to their basic framework.

Our main contribution to this literature is to provide a text-based measure of innovation that operates directly at the unit rather than patent level. That is, rather than identifying individual patents that are novel in their use of new and latterly important words we focus on the evolution of a firm, inventor or geographical area's overall patent portfolio. We are also unaware of any work on the economic modelling of innovation that has been rooted in the Bayesian Surprise concept, which has shown much utility in applications related to cognitive science (Itti and Baldi, 2009), cultural evolution (Barron et al., 2018), and the history of scientific thought (Murdock, Allen, and Dedeo, 2017).

The remainder is organised as follows. Section 2.2 introduces the methodology of our exploration measures. Section 2.3 describes the construction of our data set. Section 2.4 discusses the IBM case study. Section 2.5 applies our measures to the data and presents our main results. Section 2.6 concludes.

2.2 Measuring Exploration

To identify exploration and exploitation patterns, we first reduce the dimensionality of the data by describing the patent texts in terms of their latent topic structure. To this end, we rely on Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) – a hierarchical Bayesian model for discrete data.

In general, our approach can be summarised as follows. We start by aggregating the patent texts to documents at our desired level of analysis. This can be, for

example, at the firm-year level or represent other units of interest such as geographical regions, inventors, industries, or technology classes. We then probabilistically represent the position of each unit in the the latent topic space. The topic space can be constructed for the unit-specific sub-corpus or the entire corpus of documents. Changes in the topic shares can subsequently be measured using the concept of Bayesian surprise.

The rest of this section discusses the methodology of our exploration measures in greater detail. Section 2.2.1 describes Latent Dirichlet Allocation. Section 2.2.2 discusses Bayesian Surprise. This is followed by the definition of the measures and a discussion of their properties in Section 2.2.3.

2.2.1 Latent Dirichlet Allocation

LDA is a probabilistic topic model. The generative process described by LDA assumes that a document is constructed as a mixture of topics. As such, LDA belongs to the class of mixed-membership models that attach multiple rather than a single class to each observation.

For each document, the mixed-membership property is expressed in terms of a probability distribution over latent topics. The topics are defined as probability vectors over all words forming the vocabulary, that is, each entry represents the weight a topic assigns to the corresponding term. In this way, a topic is characterised by the probability mass it places on a set of words expressing a common theme. Note that a word can be used to represent multiple topics with different probabilities. Intuitively, in our application to patent texts, a topic represents a technology.

The advantage of LDA over other natural language processing techniques is that the generative model provides a complete probabilistic interpretation. This allows to empirically compute information-theoretic quantities based on the inferred probability distributions. Specifically, the topic distribution represents a source sending a signal – the stream of words forming the document.

To generate a set of observed documents, LDA is formally specified in terms of the following process:

(1) For each document d:

   a. Draw topic proportions $\theta_d|\alpha \sim \text{Dir}(\alpha)$.

   b. For each word $w_{d,n}$:

      i. Draw assignment $z_{d,n}|\theta_d \sim \text{Mult}(\theta_d)$.

      ii. Draw word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.

where K specifies the number of topics, $\beta_{1:K}$ are the topic specific word distributions over the vocabulary, and α is a K-dimensional Dirichlet parameter. θ_d represents the

topic proportions, z_d denotes the topic assignments, and w_d are the observed words for the d-th document.

For a given collection of documents, the inferential problem is to compute the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)},$$

where $\boldsymbol{\theta}$, $\boldsymbol{z}$, and $\boldsymbol{w}$ denote the corpus-level sets of the respective document parameters. This posterior distribution is intractable. There are several procedures to estimate the parameters including both sampling and approximation-based algorithms. Since the patent corpora we analyse are potentially very large, we appeal to variational methods to perform posterior inference. We outline the approximate posterior inference procedure in Appendix 2.B.

Model Selection

LDA belongs to the class of unsupervised learning algorithms. As such, the fundamental parameter to be prespecified when applying LDA is the number of topics K. In particular, there is a trade-off between a smaller number of topics leading to better human interpretability and a larger number of topics improving statistical measures of model-fit (Chang, Boyd-Graber, Gerrish, Wang, and Blei, 2009).

In our application to firm patent texts, we estimate individual topic models for each firm in the data set. Hence, searching for the optimal number of topics for each firm corpus is computationally expensive. Additionally, our focus lies on computing a summary measure of changes in the topic distributions rather than interpreting individual topics. For these reasons, we employ the following heuristic to set the number of topics depending on the size of the firm corpus. If a corpus of documents is comprised of more than 100 patents we set the number of topics to 50, for more than 1,000 patents to 100, and for more than 10,000 patents to 150. For corpora consisting of fewer than 100 patents we use ten topics. When estimating common topic spaces in the case of our county-level and inventor-level analysis, we set the number of topics to 100. Our results are robust to choosing different numbers of topics.

2.2.2 Bayesian Surprise

The concept of Bayesian Surprise by Itti and Baldi (2009) is the second key ingredient to the definition of our exploration measures. On an abstract level, Bayesian Surprise is a measure of how data affects an observer and is rooted in information theory and Bayesian decision theory. The underlying principles are as follows.

First, the presence of uncertainty is a necessary condition for surprise to exist. Second, surprise represents a relative deviation from an observer's expectations.

For instance, an observer may experience varying amounts of surprise at different points in time for the same data. Third, in a Bayesian framework, uncertainty is represented by probabilities that capture subjective degrees of beliefs. As data is acquired, the beliefs are updated from prior beliefs to posterior beliefs using Bayes' Theorem.

Building on these principles, Itti and Baldi (2009) define Bayesian Surprise as the difference between an observer's prior and posterior beliefs. Thus, only data which substantially affect the observer's beliefs yields surprise. They note that this is independent of the informativeness of the observation as measured by Shannon entropy, that is, the general uncertainty around the random variable's outcome.

Formally, Bayesian Surprise is computed as the Kullback-Leibler (KL) divergence from a prior distribution q to posterior distribution p

$$D_{\mathrm{KL}}(p||q) \; = \; \sum_{i=1}^{N} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}.$$

Rewriting the above equation yields

$$D_{\mathrm{KL}}(p||q) \; = \; \sum_{i=1}^{N} p(x_i) \left[\log_2 p(x_i) - \log_2 q(x_i) \right],$$

that is, Bayesian Surprise is equivalent to the expectation of the logarithmic difference between the prior q and the posterior p where the expectation is taken with respect to p. Note that when using a logarithm with base two, it is measured in bits. Also note that Bayesian surprise is asymmetric but invariant with respect to reparameterisations due to relying on the KL divergence.

2.2.3 Exploration Measures

Exploration

Based on the above definition of LDA and Bayesian Surprise, we construct our exploration measure. In general, we measure exploration from the perspective of an observer learning about the new patents applications in a given year. The observer's prior belief is the cumulative average topic distribution up to year t. That is, the observer expects the same average topics as observed in the past – exploitation is the expected default behaviour. We then measure exploration as the surprise the observer experiences when upon learning the topic distribution in year t. Put differently, we measure exploration as the temporary deviation from the past topic mean. This allows us to distinguish between phases of exploration and exploitation.

Formally, similar to the study by Murdock, Allen, and Dedeo (2017) in a cognitive sciences context, we define exploration as

$$\eta_t \; := \; D_{\mathrm{KL}}\left(\theta_t \middle|\middle| \bar{\theta}_{-t}\right),$$

where

$$\bar{\theta}_{-t} \;=\; \frac{1}{t-1}\sum_{j=1}^{t-1}\theta_j$$

denotes the average topic distribution up until year t. In our applications, the topic distribution θ_t in a given year t is based on the collection of all documents filed by the firm, an inventor or in a given county that year.

Properties

We now interpret the technical properties of our exploration measure. The mechanics are the same for all three levels of aggregation in our empirical analysis, that is, inventors, firms and counties. First, note that in the case where the observed unit's topic distribution is exactly the same as the past average topic distribution, our exploration measure is equal to zero. This corresponds to a year in which they exploit accumulated knowledge. In case it is different from the past average, our measure is greater than zero. This corresponds to a year in which they explore new topics. Additionally, note that based on the construction of using the past average as a prior, the first time an inventor, firm or county explores a new topic, our measure will be higher compared to a situation where they pick up a topic it has already worked on in the past. Hence, our measure can be interpreted as temporal novelty.

Second, as pointed out above, exploration is asymmetric due to relying on the KL divergence. As a result, it has the desirable property of attaching higher weights in situations where the share of a topic increases compared to the opposite situation where a firm works less on a specific topic compared to the past average. Therefore, our measure not only measures the difference between the current and past distributions but it also takes into account their order. This property naturally corresponds to the definition of an exploration measure.

Cumulative Exploration

In addition to the above exploration flow measure, we are also interested in characterising the life-cycle of a firm in terms of different phases of exploration. That is, we not only address the question of how surprised an observer is in a given year but we also analyse the accumulated surprise an observer has experienced following the patenting activities by the firm in the past. For this purpose, we define the cumulative exploration or 'exploration stock' in a given year t as the monotonically increasing function

$$H_t \;:=\; \sum_{j=1}^{t}\eta_j.$$

Successful Exploration

The flow and stock exploration measures allow us to quantify exploration in terms of deviations from the past topic mean. They do not, however, distinguish between successful and unsuccessful exploration. To identify phases of successful exploration, we adopt the resonance measure proposed by Barron et al. (2018). Resonance modifies the exploration measure by including a term that captures the future impact of new technologies.

In our application, this allows us to quantify the surprise of the patent topics in a particular year compared to the patterns of previous years and subtract the difference to future topics. High surprise given the past as a prior represents the unit-level exploration of new topics. High surprise given the future as a prior indicates that the unit does not continue working on the same set of topics. Hence, by considering both the initial novelty of the patents filed in a given year and the similarity to future patents, successful exploration is conceptually related to an innovation measure.

Formally, the measure is defined as follows

$$\rho_w(t) \;:=\; \frac{1}{w}\sum_{d=1}^{w}\Big[\mathrm{KL}\big(\theta_t||\theta_{t-d}\big) - \mathrm{KL}\big(\theta_t||\theta_{t+d}\big)\Big],$$

where w is the window size. To put this in words, the measure uses the KL divergence to compare the topic distribution of year t to year $t-d$. From this, it then subtracts the KL divergence between year t to year $t+d$. The differences are averaged over all years that fall into a predefined window of size w around year t. Hence, the first term in the resonance measure corresponds to the novelty of the patents in a given year, while the second term captures whether a firm works on these topics in the future.

The resulting mechanics can be summarised as follows. Resonance in a given year t is low if the technologies are similarly different from past and future technologies or very similar to both. As a result, the measure is either equal to or close to zero. Positive resonance corresponds to situations where the current technologies are different from the past average and similar to subsequent technologies. This surprise asymmetry can be interpreted as successful exploration. Note that by construction, the two terms in the resonance measure are not symmetric. This is because the second term uses the future topic distribution as a prior.

One obvious drawback of using resonance in our application is that we require future information. Therefore, while we are able to identify historic phases of successful exploration, it cannot be used in a predictive way but rather complements our analysis.

2.3 Data

This section describes the construction of our data set from several sources relating to the three levels of aggregation in our analysis, that is, inventors, firms and counties. For a more detailed description of our text pre-processing steps we refer the reader to Appendix 2.A.

Patent Abstracts

Similar to the analysis in Bergeaud, Potiron, and Raimbault (2017), we rely on patent abstracts rather than the full texts. The abstract should include the most important words that characterise the invention. Furthermore, the patent abstract focuses on the invention itself rather than including, for example, legal text.[1]

We obtain the abstracts from two main sources. Firstly, Bergeaud, Potiron, and Raimbault (2017) provide a database of abstracts for four million granted patents covering the period from 1975-2014. This database is derived from the electronic text patent records published by the USPTO. Directly inputted electronic records are not available before 1975 so we draw on a second database from Iaria, Schwarz, and Waldinger (2018). Their database was assembled from Google Patents files that were originally built up by applying optical character recognition (OCR) tools to scans of the original pre-1975 patent texts.

Formal abstract sections in patent text only became standard from the late 1960s onwards so we construct 'pseudo-abstracts' for this earlier period by subsetting the first 250 words of the patent document. This obviously relies on the assumption that the first 250 words are an effective summary of the overall patent. Our basic approach for defining the text of the abstract is to extract the text that lies between the two headings 'Abstract of the Disclosure' and 'Background of the Invention'. If the second 'Background...' heading cannot be found we define the 150 words after 'Abstract of the Disclosure' as the abstract text.

As a cross-check we compare the pooled pre- and post-1975 data to the list of 3 million patents from 1963 to 1999 that are included in the NBER legacy data set (Hall, Jaffe, and Trajtenberg, 2001). This revealed a set of 305,314 missing patents not covered by our main two datasets, so we directly webscrape information on this missing set of patents from Google Patents. The pooled dataset across the three datasets covers 7,183,108 million patents granted between 1920 and 2014. Within this total, 2,466,973 patents are represented by pseudo-abstracts.

1. While there tend to be differences in topic coherence when comparing topics based on full-text to abstract data when extracting topics from small document collections, for large document collections these differences are less significant (Syed and Spruit, 2018)

Patent Citations and Technology Classes

Our main source of data for patent citations and technology classes is the 'Comprehensive Universe of U.S. Patents' (CUSP) database constructed by Berkes (2018). In a similar vein to the abstracts, the citations are taken directly from computerized records for the post-1976 period and extracted from the text for the years prior to this. Berkes (2018) parses text from the 'References Cited' sub-section for patents issued between 1947-1975 and looks across whole body of patent texts for the pre-1947 era, focusing on keywords that suggest the quoting of explicit patent numbers.

A novel aspect of the USPTO technology class field is that the USPTO regularly updates and corrects these classifications. This means that patents can be categorised according to a consistent modern taxonomy of classes. The three main classification systems in use are the International Patent Class (IPC), the Cooperative Patent Classification CPC) and the US Patent Classes (USPC). Berkes (2018) collects the USPTO classifications as at the date of June 2016 and defines a main class based on the distribution of disaggregated 3-digit classes for the CPC and IPC, while a main class is directly identified by the USPC system.

Firm Outcomes

To connect our exploration measure to firm outcomes, we use the Compustat and CRSP databases. Compustat contains information on listed company accounts from 1950 onwards while CRSP provides us with much more limited information based around stock prices and market value back to 1925.

We use the match of patent numbers to the CRSP 'permno' identifier from Kogan, Papanikolaou, Seru, and Stoffman (2017) to connect the two sides of the data. The Kogan et al. (2017) data provides information on 7,536 firms that are matched to 1.9 million patents from 1868 - 2009, although the years before 1920 and after 2005 are sparse due to censoring. For simplicity, we only use firms with a unique mapping of permno to gvkey as found in the CRSP crosswalk file, leading to a sample of 6,544 firms matched to the patent data.

Final Firm Data Set

Our successful exploration measure depends on a 'rolling window' structure whereby current period t topic distributions are compared to past and future distributions. This creates the restriction of requiring at least 11 years of continuous data in order to calculate firm-level exploration. In turn, our main sample is therefore a subset of 1,830 unique firms who account for 1,861,219 patents in total.

We calculate our measure of firm age from the joint firm-patenting database. That is, we infer the 'birth year' of the firm as the minimum year by permno. This captures the first year that a firm appears either in the USPTO patenting data or

in the CRSP and Compustat firm data. For example, if a firm has taken out patents before it lists on the stock market, we are able to infer its existence on that basis. We finally drop the data from 2004 onwards to adjust for censoring effects such as the drop-off in patenting due to the lag between application and granting.

Geographical Data

The construction of the data sample for our county-level analysis is based on data set described above. We combine this with information on the assignee county and United States Patent Classification (USPC) classes provided by Berkes (2018). We then merge in the classification of USPC patent classes into technological categories and sub-categories following Hall, Jaffe, and Trajtenberg (2001). We obtain the mapping for this from Acemoglu, Akcigit, and Kerr (2016).

Lastly, we combine the annual exploration measure the population counts for each county from Manson, Schroeder, Van Riper, Kugler, and Ruggles (2020). Since the official population numbers are only available every five years, we linearly interpolate the population growth for the remaining years.

Inventor Age Data

For our inventor-level analysis, we obtain individual inventor identifiers and birth years for patents granted between 1976 and 2018 from Kaltenberg, Jaffe, and Lachman (2021). Their inventor birth years are inferred from information about inventors (name and location) combined with age information from different publicly available online web directories. We first merge this data with our full patent abstract sample. We then calculate the inventor ages as the difference between the application year of a patent and the birth year of the inventor. The resulting sample contains 3,264,210 patent texts matched to 1,354,897 individual inventors.

2.4 Case Study: International Business Machines (IBM) Corporation

To demonstrate our methodology, we first develop the case study of a single, long-lived firm. Specifically, we focus on the International Business Machines (IBM) corporation. IBM first emerged as a single corporation in the early 1920s from the merger of several previous companies with histories that go back to the 1880s. The company also had a central role in the development of computing technology in the 20th century, making it a good general example of the process of technological change.

We start by investigating changes in the raw word frequencies. In particular, we compute the change in the shares of a single word stem (unigram) in the total

Table 2.1. Fastest Growing Unigrams by Decade for IBM.

Overall		**1930s**		**1940s**		**1950s**	
Word	Share	Word	Change	Word	Change	Word	Change
data	2.59	mean	1.64	card	2.81	circuit	2.52
system	1.45	feed	0.85	machin	1.68	magnet	1.63
layer	1.26	select	0.61	tape	1.10	memori	1.38
first	1.23	new	0.58	perfor	0.97	data	1.19
devic	1.13	gear	0.58	electron	0.69	signal	0.94
circuit	1.02	sheet	0.55	number	0.61	input	0.90
signal	0.94	time	0.55	sens	0.56	puls	0.87
second	0.92	applic	0.47	column	0.47	line	0.77
memori	0.84	charact	0.46	digit	0.47	devic	0.76
control	0.76	invent	0.43	valu	0.46	binari	0.63
1960s		**1970s**		**1980s**		**1990s**	
Word	Change	Word	Change	Word	Change	Word	Change
surfac	0.73	silicon	0.85	data	1.18	user	0.73
cell	0.60	line	0.78	system	1.04	layer	0.59
metal	0.58	layer	0.72	imag	0.53	system	0.56
control	0.55	print	0.55	comput	0.52	first	0.40
substrat	0.54	address	0.52	first	0.49	one	0.37
code	0.50	data	0.52	document	0.44	content	0.36
error	0.46	chip	0.50	access	0.42	request	0.34
wave	0.35	region	0.50	user	0.38	method	0.32
member	0.34	generat	0.40	circuit	0.35	process	0.31
mean	0.34	ribbon	0.38	optic	0.34	inform	0.30

Notes: This table shows the fastest growing unigrams (single words) per decade. This is calculated as the change in the share of the word in the total frequency count of words used in IBM patents. We construct this from a panel of the top 500 words per year for IBM's patents. The first panel shows the top words across all years measure in terms of the levels rather than changes in share. The units are percentage points (for example: 1.64 is 1.64%).

frequency counts used in IBM patents. This is constructed as a panel of the top 500 words per year for IBM's patents. The first column in Table 2.1 shows the top words across all years measured in terms of the levels. Unsurprisingly, the word "data" has the largest overall share. The remaining columns show the fastest growing unigrams calculated as the change in the share of the word in the total frequency count of words used in IBM patents per decade from the 1930s to the 1990s.

The table illustrates the shift in IBM's technologies over time. The early periods show IBM's focus on analogue apparatuses such as punched-card machines evidenced the use of words such as "gear", "time" and "sheet" in the 1930s and "card", "machin", and "tape" in the 1940s. For example, IBM managed the administrative

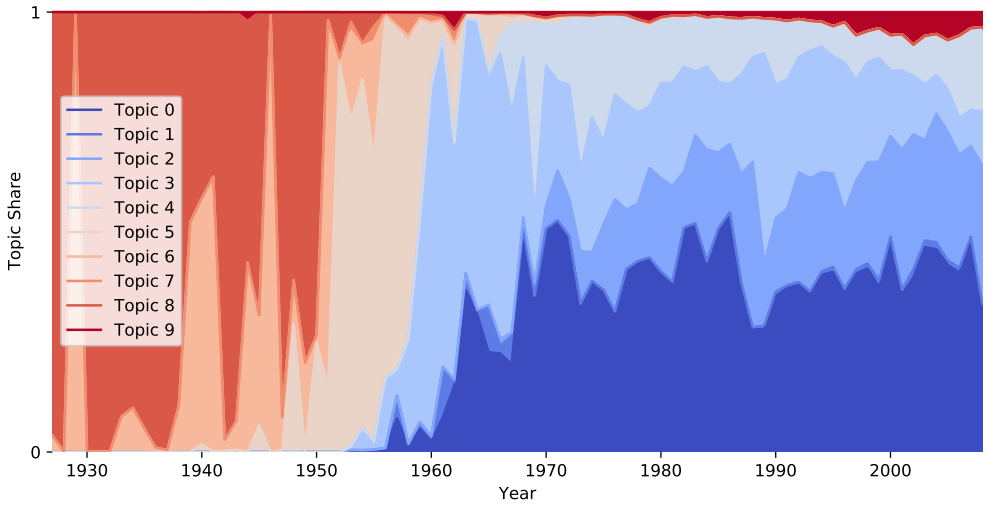


Figure 2.1. Evolution of Topic Shares for IBM.

Note: This figure illustrates the evolution of topic shares obtained from running a ten-topic LDA for IBM patents from 1927 to 2004.

information for the 26 million employment records that needed to be kept as part of the New Deal's Social Security Act of 1935.

The 1950s mark the transition from punched-card storage to digital storage (Bradshaw and Schroeder, 2003). This shift is evidenced by increases in count frequencies of words such as "circuit", "magnet", "memori", "data", and "signal". The 1960s to 1990s are characterised by words such as "surfac", "silicon", "data" and "user", respectively, representing the consolidation of the personal computer and the beginning of the internet.

Note that the growth rates after the 1950s are significantly smaller in magnitude compared to the previous period indicating that IBM stopped exploring and creating radically different inventions during this time but rather slowly adopted new technologies. This coincides with the period that lead up to the 'near-death' of the company in the mid-1990s.

We now illustrate how these changes observed at the high-dimensional word frequency level translate to the lower-dimensional topic space. First, to be able to visualise the evolution of topic shares, we run a separate ten-topic LDA model for IBM patents from 1927 to 2004 rather than the fully-fledged 150 topic specification. Figure 2.1 shows the evolution of the inferred topic shares over time. Most prominently, the graph illustrates the shift in the shares from analogue topics to digital topics in the 1950s. Furthermore, the periods before and after this transition are characterised by distinctive patterns. During the analogue era, IBM's topic shares are rather volatile implying that the attention given to individual topics is subject to

rapid shifts. The digital period is marked by more equally distributed topic shares and generally less volatility.

Next, we show how our exploration measure summarises this information. Figure 2.2 shows the exploration, cumulative exploration and successful exploration time series for IBM from 1927 to 2004 based on the topics from the full 150-topics model. The exploration graph in Figure 2.2a exhibits clear phases of exploration and exploitation which correspond to the illustration of the topic share evolution for the ten-topic LDA model. Obviously, the largest spike in exploration corresponds to the aforementioned shift from analogue to digital technologies. IBM's early growth period up until the 1950s is characterised by higher exploration volatility capturing the radical shifts in topic attention described above. Starting from the 1960s, exploration is less volatile and smaller in magnitude which can be interpreted as a long phase of exploiting the previously developed technologies. Figure 2.2b visualises the corresponding accumulation of exploration over time. Naturally, the spike in the 1950s leads to a clear bump in cumulative exploration.

Lastly, Figure 2.2c displays the successful exploration as measured by resonance. The overall graph exhibits a very similar shape as the exploration series. The 1930s show a large spike in successful exploration. As before, the 1940s are characterised by high volatility, including negative spikes. That is, during this time IBM worked on topics that they dropped in future years. Note that 1940 marks the overall minimum of the series. As before, the 1950s show a large increase in successful exploration – the transition from analogue to digital storage. After a small positive bump in the 1970s, the graph stays flat around the zero line representing a long period without significant innovations having a lasting impact.

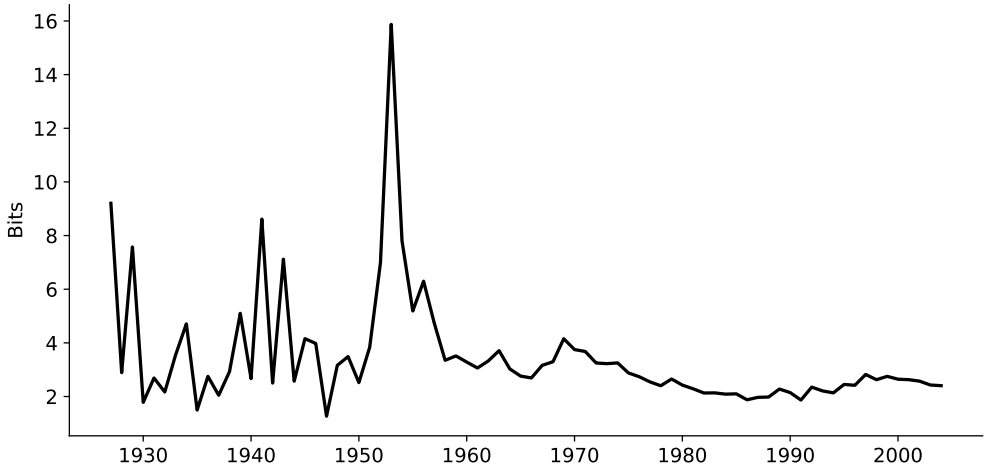
2.5 Empirical Results

This section applies our measure to the data set from the previous section and presents our main results. Section 2.5.1 investigates exploration patterns in firm behaviour and connects our measure to firm outcomes. Section 2.5.2 examines how exploration in ICT is distributed across counties. Section 2.5.3 investigates the relationship between exploration and inventor age.

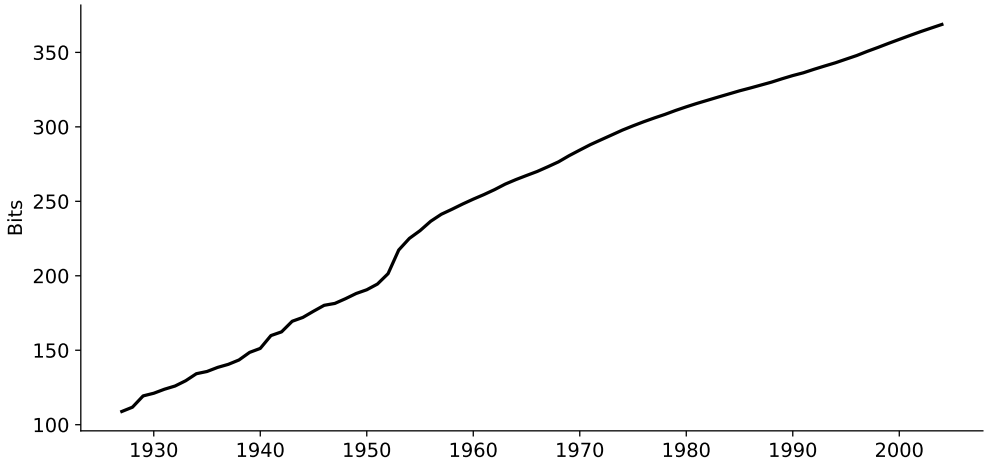
2.5.1 Exploration, Firm Age, Firm Size, and Firm Growth

Firm Age and Lifetimes

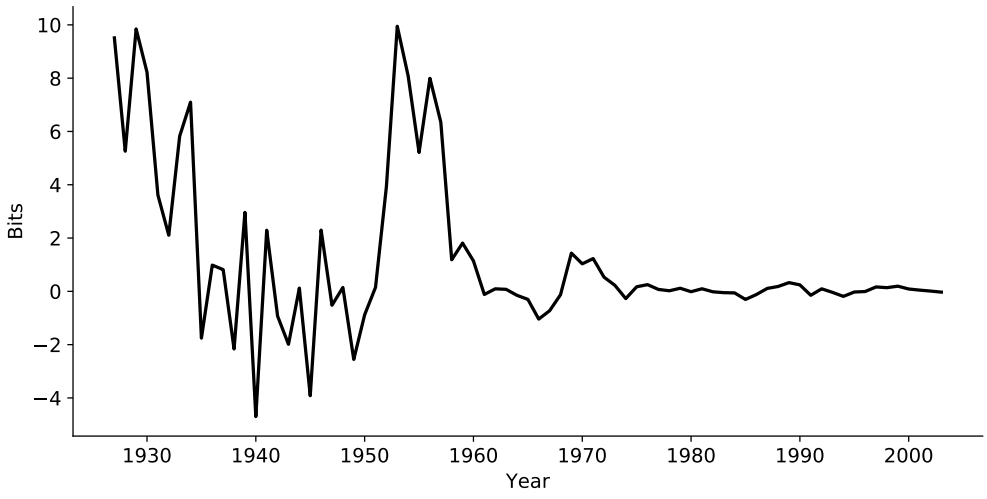
We start our analysis by presenting some information on firm ages and 'lifetimes'. Figure 2.3a shows the distribution of firm birth years amongst unique firms in the cross-section. As discussed, this is calculated as the first year a firm appears in our joint USPTO-CRSP-Compustat database.



(a) Exploration



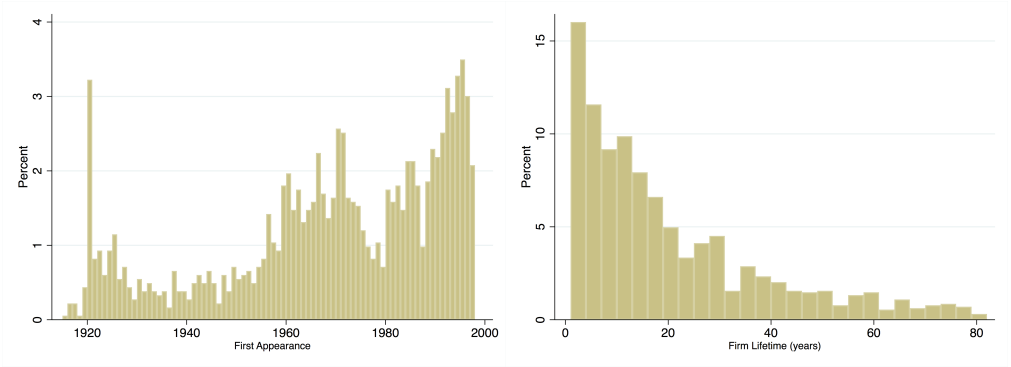
(b) Cumulative Exploration



(c) Successful Exploration

Figure 2.2. IBM's Exploration.

Notes: This figure shows the standard, cumulative and successful exploration series for IBM from 1927 to 2004.



(a) Entry Year of Firms **(b)** Firm Lifetimes

Figure 2.3. Firm Age and 'Lifetimes'.

Notes: Figure 2.3a shows the 'entry year' of the 1,830 firms in our sample. The histogram bars are defined as 1-year intervals. Figure 2.3b shows the distribution of 'lifetimes' for firms existing at or before 1990. This represents 1,286 distinct firms, with an average lifetime of 19.9 years and median of 14.

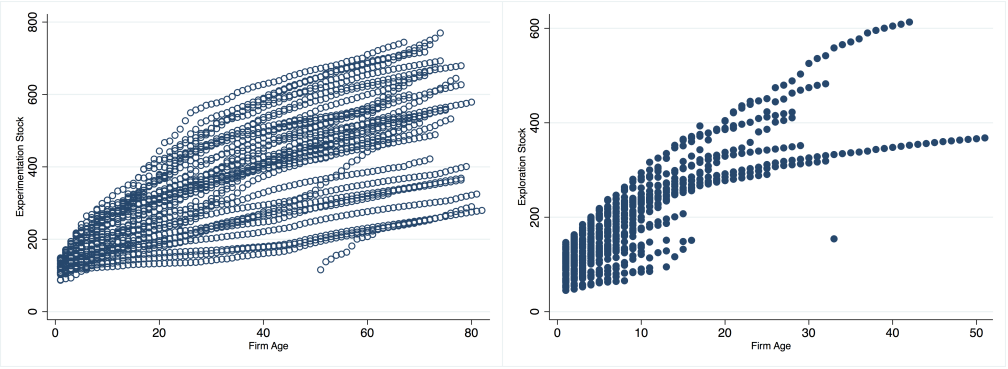
Figure 2.3b then plots a histogram on firm 'lifetimes' in the cross-section. In the computation, we consider all of the unique firms that existed before 1991 and calculate the total number of years they are contained in our data. The conditioning of the data on 1990 and before helps to account for censoring – by definition those firms that have been born recently still need time for their commercial life-cycles to play out.

Firm Topics

Before examining exploration patterns, we briefly describe the process we rely on to construct firm-year documents for the LDA inferential procedure. In a first step, we combine all patents into a single document for each firm-year. We then normalise the length of this document to 100,000 words. The main reason for this is that the normalisation helps establishing comparability between years with different numbers of patent applications. The choice of 100,000 words is robust in the following sense. While shorter documents would introduce a noticeable bias to our exploration series, for document lengths above this number, our results do not change significantly. From these documents, we then infer the firm-level topic distributions which form the basis for our exploration measures and are used throughout this section.

Trends in Exploration

How does exploration evolve over the life-cycle of long-lived forms? Figure 2.4a displays the paths of the exploration measure for a sample of large, long-lived firms – aged 60 or older by the end of the sample and included in the top 5% of firms in



(a) Large Firms **(b)** Semiconductor Industry

Figure 2.4. Exploration Stock and Firm Age Over Time.

Notes: Figure 2.4a shows the evolution of our 'experimentation stock' measure for firms that are aged 60 or more as of 2007 and are above the 95th percentile in the firm-level distribution of total cumulated patents (practically, 3,357 patents). N = 35 for the number of firms included. Average age of firms is 72.1 years. Figure 2.4b shows the evolution for firms in the semiconductor industry (SIC4=3674). N = 82 firms. Average age of firms is 10.9 years. Average cumulated number of patents per firm is 1,257.7. The SIC code assigned in Compustat from 1950 onwards is assigned for firms existing as part of the CRSP data pre-1950.

terms of total patents. These paths show evidences of clearly defined trends at the firm-level, including indications of classic 'S-shaped' developmental behaviour.

We follow this up in Figure 2.4b by conditioning on all firms in the semiconductor industry but relaxing any constraints on minimum firm age. This shows a pattern of dispersion whereby firms with higher exploration trajectories appearing to 'breakaway' after surviving their first 10 years.

Next, we turn to regression models to further investigate this relationship. In particular, we aim at disentangling the question of how exploration varies with age and whether this relationship is conflated with firm size. We use the cumulative exploration measure or 'exploration stock' as the dependent variable. Table 2.2 shows the results for different specifications. The main message is that exploration is indeed parabolic in age and, interestingly, age explains exploration over and above any correlation with firm size. Specifically, Columns (3)-(5) control for market capitalisation, the firm patent stock and firm sales in succession with minimal effects on the coefficients of the two age variables. That is, age dominates as a stronger correlate of exploration, with this being clearly evident in the raw correlations. For example, the age-exploration correlation is 0.83 compared to 0.38 for (log) market cap-exploration in the data underlying the Column (3) regression.

To summarise these relationships, we plot the age-firm size and age-exploration gradients in Figure 2.5. These gradients are the predictions from pooled cross-sectional regressions of the outcomes controlling for year effects. They show that exploration has a less steep slope with respect to age, that is, exploration tapers

Table 2.2. Relationship between Cumulative Exploration and Firm Age.

	(1) Baseline	(2) +SIC4	(3) +Mktcap	(4) +PatStock	(5) +Sales
age	12.03***	12.31***	11.89***	10.94***	11.73***
	(0.533)	(0.505)	(0.511)	(0.574)	(0.576)
age2	-0.0645***	-0.0644***	-0.0633***	-0.0562***	-0.0607***
	(0.00781)	(0.00742)	(0.00735)	(0.00807)	(0.00803)
log marketcap			7.156***		
			(1.746)		
log patstock				14.09***	
				(2.238)	
log sales					7.890***
					(2.001)
R-sq	0.620	0.718	0.720	0.728	0.726
N	26,727	26,721	26,375	26,721	23,009

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of the cumulative exploration measure on firm age – age is the linear term while age2 is the quadratic. log marketcap is the logarithm of market capitalization, log patstock is the logarithm of the patent stock and log sales is the logarithm of sales. Year effects in all regressions, SIC4 fixed effects from Column (2) onwards.

faster with age than with firm size. Theoretically, this is interesting insofar that it shows that firm growth continues after exploration has attenuated, hinting at the existence of major phases of exploitation activity amongst firms.

There is also a clear relationship between firm age and R&D intensity (defined as R&D expenditure divided by sales), which we plot in Figure 2.6. The graph shows that R&D intensity falls with age right up until age 40. Average R&D intensity in the early years of firm lifetimes is around 0.102 (i.e. R&D spending is 10.2% of sales) with a sample mean of 0.056. Again, this is prima facie evidence of intense exploratory behaviour earlier in firm life cycles.

Exploration and Firm Growth

We now connect our exploration measure to firm outcomes in a regression framework. We look at both the short-run dynamics of exploration and firm sales (effectively 1-year growth models) as well as medium-run relationships (5-year growth models). The basic model that we adopt is as follows:

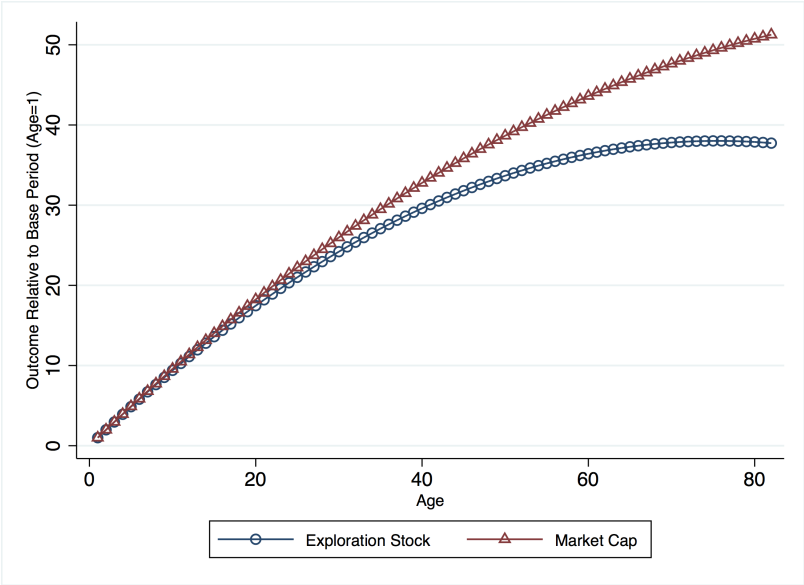


Figure 2.5. Gradients of Exploration Stock and Firm Size with Firm Age (All Firms).

Note: This figure shows the gradients of the exploration stock and firm size (defined as market cap) with firm age. This is defined as the predictions from a pooled cross-sectional regression of the outcomes on age and age squared with controls for year effects. N = 27,760 observations in the regression covering 1,795 distinct firms. The y-axis shows the level of the outcomes with respect to the age = 1 base period (i.e. we normalise with respect to initial values).

$$\Delta_k \ln(\text{Sales})_{ijt} = \alpha + \sum_L \beta_{k-1} \text{KL}_{t-l} + \tau_t + \mu_j + \tau_{jt} + \varepsilon_{ijt}$$

where $\Delta_k \ln(\text{Sales})_{ijt}$ is the k-year change in firm i log sales measure in period t, KL_{t-l} is an l-period lagged exploration measure, τ_t are time effects, μ_j are industry effects, τ_{jt} are industry trends, and ε_{ijt} is an error term. We use different lag orders L to understand the dynamic relationship across specifications.

The main model that we focus on here is the 5-year changes model. This specification is useful for 'smoothing out' variation and reducing measurement error. In Figure 2.7 we present results for a specification that uses the 5-year change in (log) sales as the dependent variable and includes single-year exploration measures on the right-hand side. In effect, this is measuring the association between a 1-year shock in exploration at $(t-k)$ on a smoothed, 5-year measure of firm growth.

Figure 2.7(a) indicates that exploration has a medium-run association with sales growth. A positive association becomes evident at around the $(t-9)$ or $(t-10)$ lags, but is quite persistent once this point is reached. Note that this specification is run in changes and uses 'flow' measures in exploration so it is differencing out fixed unobservables at the firm-level. Figure 2.7(b) then runs a similar specification but uses successful exploration as the explanatory variables of interest. This shows a much sharper, short-run effect starting at the $(t-6)$ lag and is compatible with

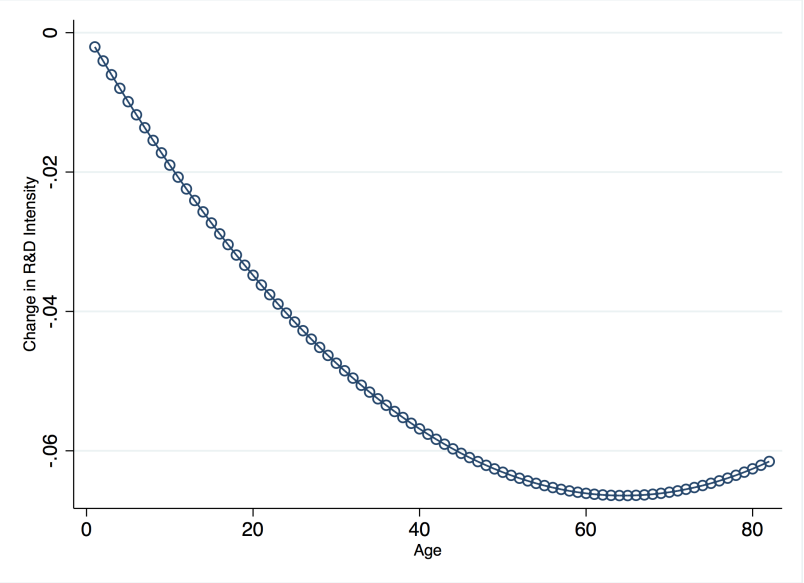


Figure 2.6. Change in R&D Intensity with Firm Age (All Available Firms).

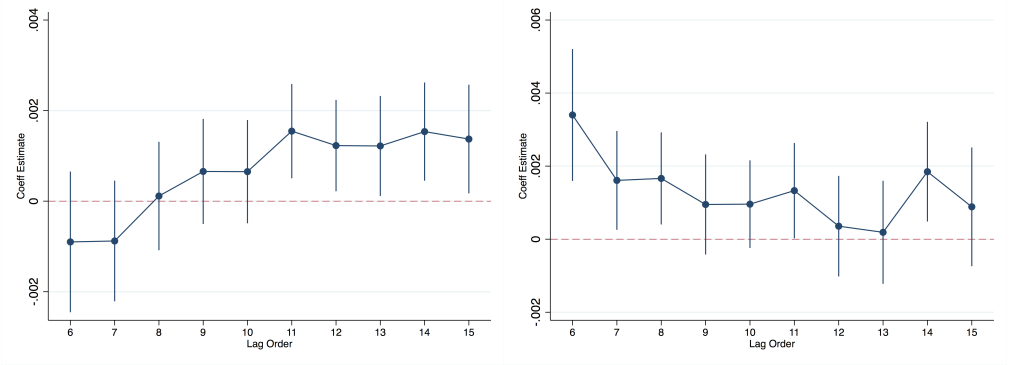
Note: This figure shows the gradient of firm R&D intensity (define as R&D expenditure over sales and firm age. This is defined as the predictions from a pooled cross-sectional regression of the R&D intensity on age and age squared with controls for year effects. N = 16,209 observations in the regression covering 1,467 distinct firms. The y-axis reports how R&D intensity changes with age. The mean R&D intensity across the sample is 0.056 while the mean starting value (i.e. at age=1) is 0.102.

the idea that the successful exploration measure is better at picking out the most effective episodes of exploration.

In Appendix 2.C we present the results for a range of alternative specifications that relate sales to exploration. In Table 2.C.1 we look at the relationship in terms of contemporaneous 1-year changes. This again shows a positive association that holds even after controlling for 4-digit industry trends, firm age and the change in the volume of patenting. The point estimate for successful exploration measure is also around three times higher than that for the standard exploration measure, confirming its effectiveness. We present the results of a similar 5-year changes specification in Table 2.C.2. This differs from Figure 2.7 by using 5-year averages of exploration on the right-hand side and confirms the same patterns as the 1-year estimates.

What is the magnitude of this association? Our exploration measures are defined in terms of information 'bits'. Hence, for example, a 1-bit increase in exploration corresponds to an (approximate) 0.1 percent increase in sales in the specification in Column(3) of the upper panel in Table 2.C.1. A 7.1 bit increase in exploration (which is equivalent to the standard deviation for this sample) then corresponds to a 0.9 percent increase in sales.

Recall here that the 'bits' are effectively measuring the extent of the *change in text information* in the firm-level patent portfolio. The regression specifications therefore



(a) Ten years of Lagged Exploration **(b)** Ten years of Lagged Successful Exploration

Figure 2.7. Five-Year Changes in Sales and Lagged Exploration.

Notes: This figure shows the estimates of a regression of the 5-year log change in firm sales on (simultaneous) lags of the general and successful exploration measures. Standard errors clustered by firm and 95% confidence intervals reported.

show that firm sales performance is correlated with this change in text information over and above the quantity of patents being produced by firms.

2.5.2 The Geography of Exploration in ICT

In this section, we investigate how exploration is distributed across space. We focus on the specific context of patenting innovation in ICT, a key driver of U.S. innovation dynamics in post-war period. We are thus interested in understanding where exploration in ICT takes place, whether it is concentrated in particular exploration hubs and what, if any, are the dynamics of the geographical distribution of exploration in ICT.

We do this in the context of an increasing polarization of economic activity across space, at least partly driven by the rise of high-tech innovation hubs in the second half of the 20th century (Moretti, 2012, 2019). Indeed, as shown by Andrews and Whalley (2021), after reaching a trough in the 1980s, the spatial concentration of patenting is today at an historical maximum, comparable to that observed in the mid-19th century. As their analysis documents, this is partly driven by the rise of ICT: by 2016, the commuting areas of San Jose (including much of Silicon Valley) and San Francisco, account for about nearly 20% of all U.S. patenting. Against this backdrop, we ask whether the spatial distribution of exploration in ICT simply reflects the patenting dominance of the familiar IT hubs or whether it is, instead, differentially concentrated.

Our ICT subsample is comprised of all patents belonging to patent category two ("Computers and Communications"). Given the focus on ICT, we further restrict the sample to patent applications made during the period from 1947 to 2007. We then

Table 2.3. Top Ten ICT Patenting and Exploration Counties.

(1) Rank	(2) County	(3) Share	(4) Rank	(5) County	(6) Share
1	Santa Clara County (CA)	31%	1	Madison County (AL)	27%
2	Westchester County (NY)	16%	2	Maricopa County (AZ)	14%
3	New York County (NY)	9%	3	Contra Costa County (CA)	9%
4	Cook County (IL)	7%	4	Alameda County (CA)	9%
5	King County (WA)	7%	5	Pima County (AZ)	9%
6	Middlesex County (MA)	6%	6	Marin County (CA)	5%
7	Harris County (TX)	5%	7	Riverside County (CA)	4%
8	Los Angeles County (CA)	5%	8	San Francisco County (CA)	4%
9	Union County (NJ)	4%	9	Orange County (CA)	3%
10	Dallas County (TX)	4%	10	San Diego County (CA)	3%

Notes: The table shows the top ten counties by shares of patenting (left) and exploration (right).

infer the topics by running LDA on the entire corpus of ICT patents aggregated at the county-year-level. This is followed by calculating the exploration measures based on the topic shares for each county. The advantage of this approach is that the topics are comparable across counties. In particular, for this exercise, we are interested in comparing the distribution and evolution of county-level exploration across the shared technology space rather than calculating within-county exploration. Hence, by using common topics, our resulting measures are not only comparable in terms their unit but also regarding the underlying topic structure.

Reflecting the highly spatially concentrated nature of patenting in ICT, the typical US county does not innovate in ICT: over the sixty-year period we consider, 2723 counties (out of a total of 3167) have zero patents, a further 285 counties patent only sparsely in ICT, with less than three patents per year on average, while the top 5% of counties account for 98% of all 452,889 ICT patents issued during this period. Henceforth we concentrate our analysis on this latter subset of counties accounting for the vast majority of ICT patenting.

Table 2.3 along with Figure 2.8 provide further confirmation of the spatial concentration of ICT patenting in the US post-war period. In particular, we compute, for each county, the total number of issued ICT patents as a share of the national grand total over the 1947-2007 period. Columns (1) to (3) of Table 2.3 rank the top ten counties while Figure 2.8a gives a heat map of the distribution across space. Consistent with our discussion above, the top ten counties account for nearly 90% of all ICT patenting during this 60 year period, with Santa Clara County alone (where Palo Alto is located) accounting for large 31% of all ICT patents and Westchester County (NY), where the IBM headquarters are located, accounting for a further 16%.

Also present in this top ten are the hubs of large metro areas (New York, Chicago's Cook County, Seattle's King County, Houston's Harris county, Los Angeles and New Jersey's Union County) as well as Middlesex County (MA), where Cambridge is located.[2] The map visualises that the counties accounting for the remaining ten percent of patenting are spread across the country with the main areas located in the East and West.

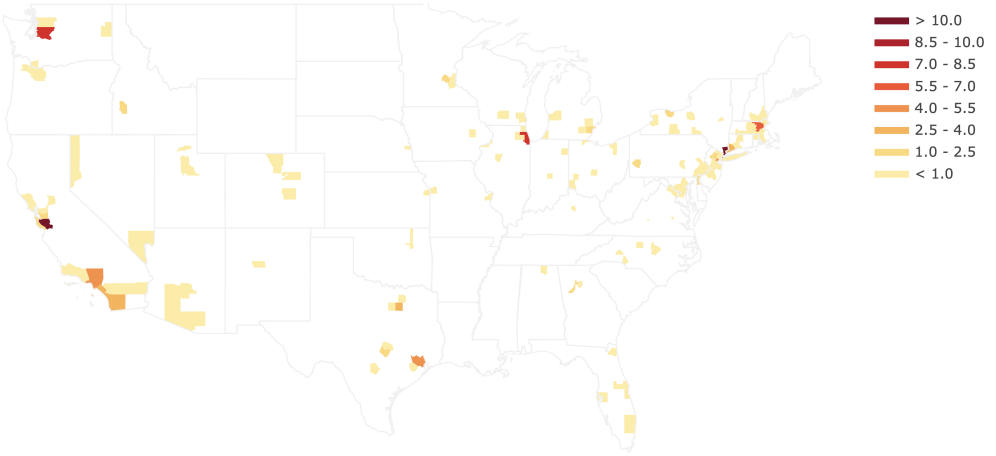
Columns (4) to (6) in Table 2.3 show the ranking of the top ten counties with the highest total exploration over the sample period. The main observations from the table are as follows. First, there is no intersection between the previous group of top ten patenting counties and the top exploring counties. This indicates that the number of ICT patents does not necessarily capture the exploratory dimension of firm innovation behaviour. This is supported by an overall rank correlation of 0.02 for all counties. Second, exploration is more concentrated at the state-level compared to patenting. In particular, nine out of the top ten ICT exploration counties are in the West, with seven located in California. Third, the exception to this previous observation is the top county Madison County (AL). The county alone accounts for 27% of the total ICT exploration. Together with the second most exploratory county Maricopa County (AZ), the top two counties represent 41% of exploration. Figure 2.8b displays the corresponding map illustrating that exploration is highly concentrated in the top ten counties that represent 87% of the total ICT exploration. We also observe the general concentration in the West in contrast to patenting.

To get a better understanding of the firms that drive the patenting exploration in the top ten counties. Table 2.4 shows the top five patenting firms for each county in the top ten. Table 2.5 shows the top five exploring firms for each county in the top ten.

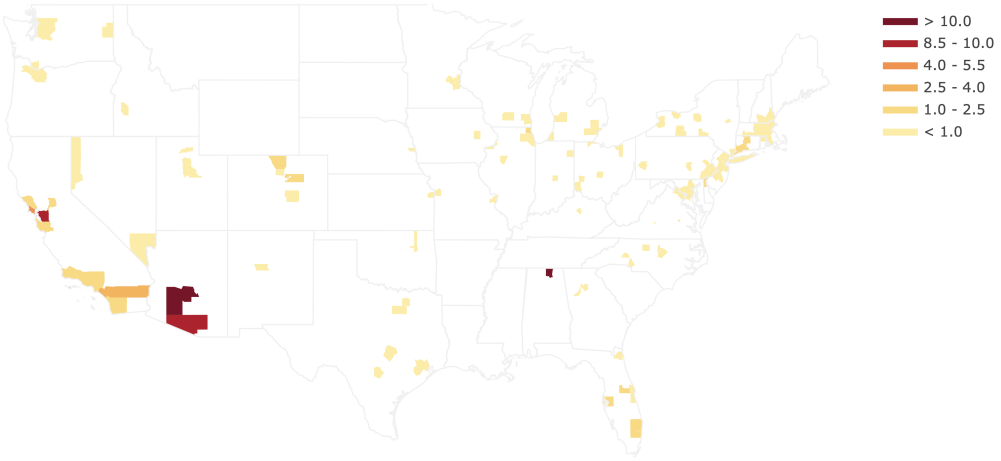
Focusing first on patenting, and in particular in the top patenting firms present in the very top three counties (which account for more than half of all patents issued over the entire period), we recognize that ICT patenting in these counties is - perhaps not surprisingly - dominated by well-recognized computer hardware component manufacturers, such as Intel, Sun, HP, Cisco, IBM (across two locations) or Hitachi as well as communications devices and services firms, such as ATT, or Phillips and an older cohort of firms in the same sector, such as ITT, RCA or Dictaphone.

Interestingly, and consistently with the limited overlap between top patenting and top exploration counties, the firms appearing as top explorers in the top exploration counties are in general distinct. For example, Madison county, responsible for more than a quarter of all ICT exploration over this sixty-year period, is a ma-

2. These findings, both regarding the scale of concentration and the identity of the particular top locations, are consistent with the patterns documented in Andrews and Whalley (2021), albeit specialized here to ICT.



(a) Top Patenting County Shares.



(b) Top Exploration County Shares.

Figure 2.8. Top ICT Patenting and Exploration Counties.

Notes: The figure shows the total number of issued ICT patents as a share of the national grand total over the 1947-2007 period.

jor aerospace and defense industry hub. The U.S. Space and Rocket Center, NASA's Marshall Space Flight Center, and the United States Army Aviation and Missile Command are all located in this county. Thus, Madison's top exploration location reflects the presence of major contractors in the aerospace and defense sector, such as Intergraph (an early developer of geographical information systems for real time missile guidance purposes), the Aviation Corporation's Research Laboratory (Avco) or SCI Systems, a major electronic component manufacturer for the defense industry, as well as communications networks firms like Motorola and Adtran. The presence of major contractors to the defense industry extends to other top exploration loca-

Table 2.4. Top ICT Patenting Firms.

Santa Clara County	Westchester County	New York County	Cook County	King County
Intel	Intl Business Machines	At&T	Motorola Solutions	Microsoft
Sun Microsystems	Hitachi	North American Philips	Boeing Co	Boeing Co
Hp	Texaco	Itt	Zenith Electronics	At&T Wireless Services
Cisco Systems	Dictaphone	Intl Business Machines	Gte	Amazon.Com
Advanced Micro Devices	Itt	Rca	At&T	Sundstrand
Middlesex County	**Harris County**	**Los Angeles County**	**Union County**	**Dallas County**
Raytheon Co	Hp	General Motors Co	Lucent Technologies	Texas Instruments
Digital Equipment	Compaq Computer	Northrop Grumman	At&T	Stmicroelectronics Nv
Emc/Ma	Litton Industries	Rockwell Automation	Alcatel-Lucent	I2 Technologies
Honeywell International	Exxon Mobil	Trw	Exxon Mobil	E-Systems
Gte	Halliburton Co	Directv	Agere Systems	Dallas Semiconductor

Notes: The table shows the top five patenting firms for the top ten counties shown in Table 2.3.

Table 2.5. Top ICT Exploring Firms.

Madison County	Maricopa County	Contra Costa County	Alameda County	Pima County
Intergraph	Honeywell International	Bio-Rad Laboratories	Network Equipment Tech	Burr-Brown
Avco	Honeywell	Chevron	Lam Research	Ventana Medical System
Motorola Solutions	General Electric Co	Systron-Donner	Exar	
Sci Systems	Motorola Solutions	Schlumberger	Eastman Kodak Co	
Adtran	Gte	Intraware	Sybase	
Marin County	**Riverside County**	**San Francisco County**	**Orange County**	**San Diego County**
Autodesk	Steris	Chevron	Western Digital	General Dynamics
Sonic Solutions	Toro Co	Macromedia	Smithkline Beckman	Cubic
L3Harris Technologies		Dolby Laboratories	Rockwell Automation	Titan
Inference -Cl A		Sharper Image	Emulex	Viasat
Fair Isaac		Schwab (Charles)	Qlogic	Oak Industries

Notes: The table shows the top five exploring firms for the top ten counties shown in Table 2.3.

tions beyond Madison county: Honeywell Aerospace and Honeywell International (in Maricopa, AZ, also a aerospace and defense hub), Systron-Donner (in Contra Costa, CA), L3Harris Tech (in Marin county), Rockwell Automation (Orange County, CA) or General Dynamics, Titon, Cubic or Viasat, all in San Diego County (CA), another major defense industry hub.

Finally, it's worth noting that beyond aerospace and defense, top explorer firms in top exploration counties reflect a diverse set of sectors, such as energy (e.g. Chevron, Schlumberger) or life sciences (e.g. Bio-Rad Laboratories, Ventana Medical, Smithkline Beecham, Steris Corp.) alongside perhaps more recognizable electronics components and devices or software firms (e.g. G.E., Autodesk, Dolby or Western Digital).

Overall, the analysis above suggests that the differential geographical distribution of ICT patenting relative to ICT exploration reflects the fact that whereas patenting is dominated by the location of electronics super-star patenting firms (such as IBM), ICT exploration reflects (*i*) innovation activities across a broader spectrum of sectors and, in particular, (*ii*) a sizeable contribution of the aerospace and defense industry, therefore tracking its geographical distribution.

The findings above suggests that, over our sample period, both ICT patenting and exploration are highly concentrated (albeit in different locations). A set of questions follow suit. Is ICT exploration more concentrated across space than ICT patenting? Are there differential dynamics of spatial concentration? Finally, how do we deal with the fact that top ranked counties according to either criteria appear to reflect very different sized counties? For example, for the year 2000, the population of Santa Clara (CA) county is close to 1.7 million while Madison County (AL) is close to 300,000. To address these questions, we follow the dartboard approach by Ellison and Glaeser (1997). The latter gives an intuitive null model to observed concentration patterns over space: that which would obtain if innovation – be it patenting or exploration – was randomly distributed across space with weights given by the population distribution across U.S. counties.

In particular, we use our data to compute the index for concentration C_t at time t as follows:

$$C_t = \frac{\sum_{i=1}^{n} (\text{Innovation Share}_{it} - \text{Population Share}_{it})^2}{1 - \sum_{i=1}^{n} \text{Population Share}_{it}^2} \tag{2.1}$$

where $\text{Innovation Share}_{it}$ is either the share of all exploration or all patents in ICT attributed to county i at time t. Whenever $C_t = 0$ this implies that each county innovation output is distributed according to its population share while if $C_t = 1$ all innovation in a given year t is attributed to a single county.[3]

3. A related alternative would be to follow a dartboard approach of exploration relative to patenting. We would then be asking whether exploration is more concentrated relative to a case

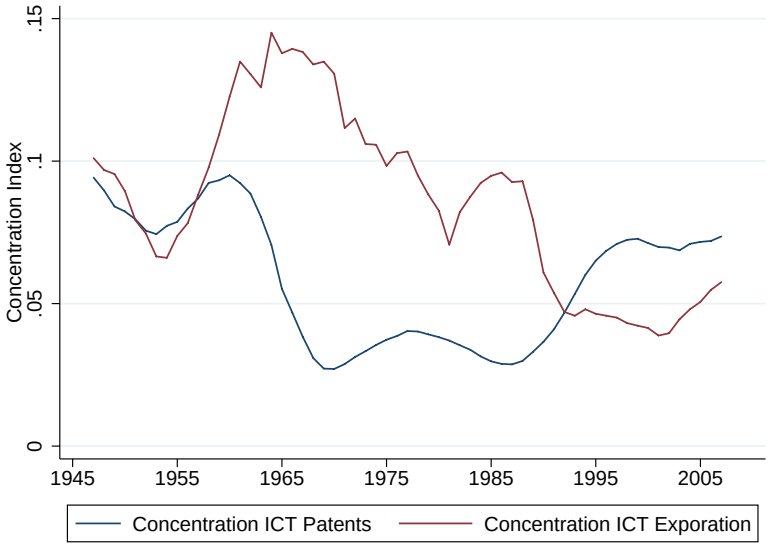
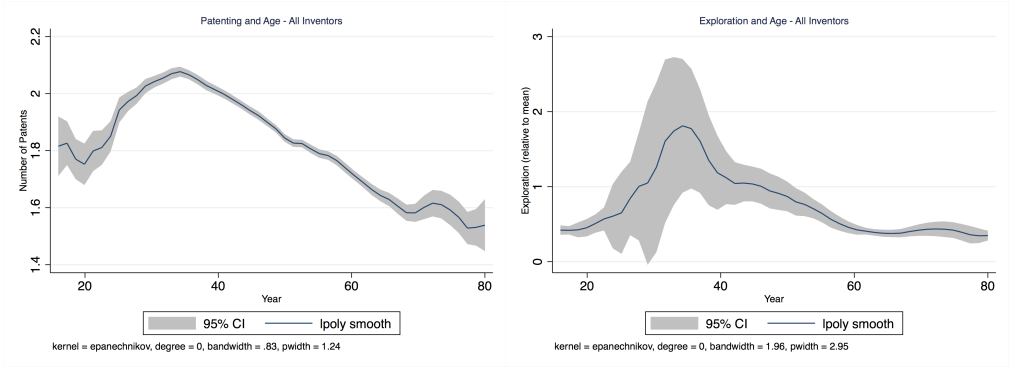


Figure 2.9. Spatial Concentration

Note: This figure shows the spatial concentration of ICT patenting and exploration from 1947 to 2007 with a five-year moving average filter applied to each series.

Figure 2.9 displays the results, where we have applied a five-year moving average filter to each series in order to focus on lower frequency movements. First, note that both series display excess spatial concentration relative to the common benchmark, the spatial distribution of population. Second, the concentration of ICT patenting displays a market U-shape pattern, with spatial concentration falling by about 50% during the 70s and 80s (relative to the 50s and early 60s) and then rising again from the mid-90s onward. Further, these ICT patenting concentration dynamics are consistent with those reported by Andrews and Whalley (2021) for the entire population of US patents. Third, the average spatial concentration of exploration in ICT is higher than that of patenting (0.09 versus 0.06 sample averages, respectively). Fourth, this is chiefly due to the different dynamics of the two time series. Thus, though they start at comparable levels of spatial concentration in the 50s, by the early 60s, when patenting concentration declines, exploration concentration increases (by about 50%) throughout that decade and, despite then initiating a trend decline, its excess concentration (relative to patenting) remains high throughout the 70s and 80s. By the same token, when we observe patenting concentration

where exploration would be distributed across U.S. counties according to their respective ICT patenting shares. Not surprisingly, and anticipating results, this alternative approach yields similar findings to those presented above: exploration is more spatially concentrated than patenting but this excess concentration has declined over the decades.



(a) Patenting. **(b)** Exploration.

Figure 2.10. Patenting and Exploration per Age for All Inventors.

Note: This figure shows local polynomial regression plots for the sample of all N = 300,561 inventors with ages between 16 and 80. Patents are allocated in full to co-invented patents. The exploration measure is an index: exploration in 'bits' divided by the sample average of exploration.

increasing again in the 90s, this is when we see exploration concentration declining below (that of patenting).

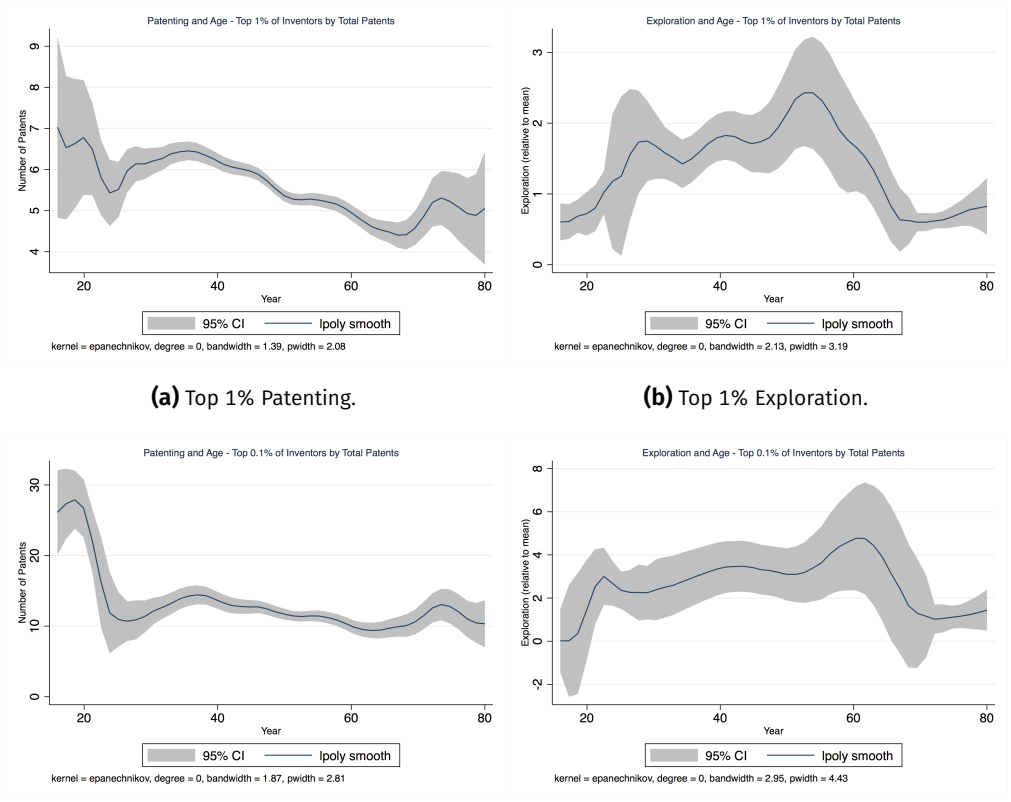
2.5.3 Exploration Over the Course of Life

In this section, we investigate the relationship between exploration and inventor age. Conceptually, our measure of exploration allows us to address the classic question of how scientific creativity varies with age. A broad range of research has suggested that creativity peaks in the age decades of the 30s and 40s. Empirically, research on this topic has been obliged to use proxy measures of creativity such as patent or publication counts weighted by citations. In contrast, our exploration measure is designed to directly track how a researcher moves through 'knowledge space' over the course of their work.

We estimate inventor-level exploration by first estimating a 100-topic LDA model across all patent documents over all years. Exploration is then defined according to an inventor's topic shares for the portfolio of patents they produce in a given year. Hence, exploration in this context can be interpreted as measuring the shift in an inventor's pattern of specialisation across a set of topics defined at the level of the population corpus.[4]

Figure 2.10 shows the results of a local polynomial regression of outcomes on age for all inventors in the sample. In panel (a) we report the age profile of patent-

4. Note that in contrast our firm-level analysis uses the firm-specific corpus to define the initial topic model, allowing for a 'within-firm' analysis of changing specialisation. We adopt the population-level corpus for inventors mainly for pragmatic (computational) reasons.



(a) Top 1% Patenting.

(b) Top 1% Exploration.

(c) Top 0.1% Patenting.

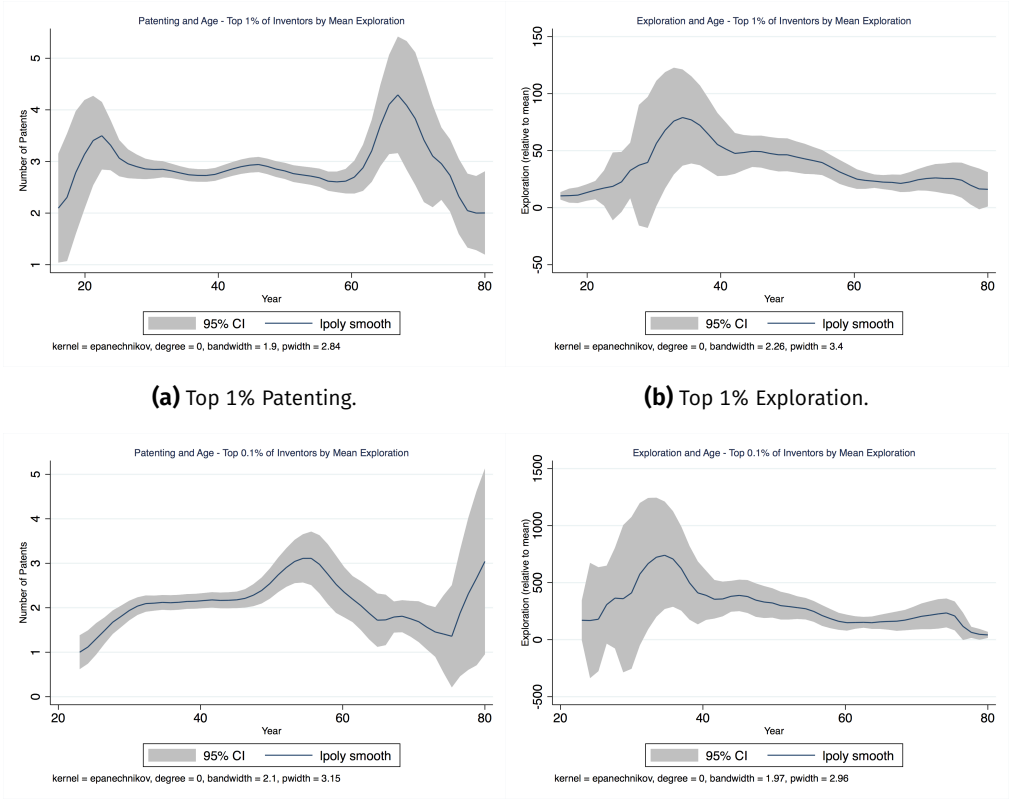
(d) Top 0.1% Exploration.

Figure 2.11. Patenting and Exploration per Age for Top Patenters.

Note: This figure shows the results of local polynomial regressions for the samples of the top 0.1% and 1% patenting inventors.

ing – effectively patenting productivity over the life-cycle. The result here directly mirrors that of Kaltenberg, Jaffe, and Lachman (2021) – productivity in terms of patenting volume peaks around the age of 40 and then declines. Panel (b) then plots the profile for exploration, where we have normalised exploration according to the sample mean such that the y-axis can be interpreted as an index. This also shows a peak at around age 40. In this case, it is a steeper peak. Exploration is 2-3 times higher in the age 30-40 range than it is at other points in the life-cycle.

How does the exploration profile evolve for the most prolific inventors? We plot the age profiles for the top 1% of inventors by the number of total patents in panels (a) and (b) of Figure 2.11 and then the top 0.1% in (c) and (d). This shows more variability in patenting productivity, with 'bursts' early and late in the life-cycle, but a high productivity mid-life phase is still evident. In terms of exploration, it should be first noted that the top 0.1% of inventors also tend to be more exploratory on average with indexed exploration levels of around 3.5-4 in mid-life compared to



(a) Top 1% Patenting.

(b) Top 1% Exploration.

(c) Top 0.1% Patenting.

(d) Top 0.1% Exploration.

Figure 2.12. Patenting and Exploration per Age for Top Explorers.

Note: This figure shows the results of local polynomial regressions for the samples of the top 0.1% and 1% exploring inventors.

1.5-2.0 for the full sample. Exploration also progresses in 'waves' across the life-cycle with a high level of exploration spread across the decades from the 30s to the late 50s.

We do an additional split by the top exploring inventors in Figure 2.12. That is, we calculate average exploration over the life-cycle and pick out the top 1% and top 0.1%. This results in a sample of inventors who produce an average of 2-3 patents per year. In this case, the pattern of exploration follows the more conventional pattern of peaking close to the age of 40 without any subsequent 'waves'. Arguably, what is most notable about this set of 'top explorers' is that an age profile is still evident even though these inventors have high baseline levels of exploration.

2.6 Conclusion

In this paper, we provide a new measure of unit-level exploration and exploitation. We empirically connect the measure to key questions in the literatures on firm

growth, inventor life-cycles and the geography of innovation. We find evidence of exploration patterns in firm behaviour that are distinct from other potentially correlated aspects of firm performance, a mid-life peak in exploration for inventors, and evidence that exploration is geographical concentrated within the US but that this is coming from the 'periphery' rather than the main hubs of patenting.

The generalisability of our results faces a set of limitations. First, patent data is inherently biased towards a given unit's exploration activity that resulted in a patent application. Hence, while we rely on patents as an imperfect proxy for the total exploration activity, it is impossible to observe all innovation efforts. In addition, we only consider granted patents and thus exclude patents applications that were rejected. Second, in the case of firms and inventors our data set is subject to survivorship bias in the sense that we focus on the units with longer histories. Therefore, it is unclear how our results carry over to newer firms or inventors. Third, we currently do not take into account the effect of strategic interaction and renewal periods on patenting activity. Fourth, similar to most applications of natural language processing to a large, historic corpus there might be underlying changes in the patent language. However, since technical language typically faces less change compared to other written or spoken language, we deem this not to be too big of an issue.

Future Work

We plan to develop the work in this paper in the following directions, with a strong focus on firms. Our first direction involves deepening the present analysis and further characterising the prevalence of exploration versus exploitation across the size and age distribution of firms. For example, to what extent do young or small firms engage more heavily in exploration? Do firms engaging in exploration grow faster, either during or, more likely, after periods of successful exploration? Are they more profitable? Are current superstar firms more likely to have engaged in exploration at some point in their life-cycle? To what extent does heightened exploration correlate with a higher intensity of more typical innovation inputs like R&D ratios or outputs, such as patent citations and firm-level measures of productivity?

As a second direction, we will extend the breadth of our text-based measures of exploration. Our current measure focuses on the variance of exploration within a unit's life-cycle and have less explanatory power for studying how a unit's innovation behaviour is different from its peers. We are currently working on the implementation of an additional measure based on the Jensen-Shannon divergence that is better targeted for quantifying firm deviations from group average. There is also scope to complement our divergence measures with simpler metrics such as those based on how important, new words enter and diffuse through the patents text corpus.

As a third direction, we plan to aggregate our firm-level measures at the industry and economy level to explore a wider range of questions: do we observe exploration at the industry level or do firm-level decisions wash out at the industry level, simply inducing innovation reallocation across firms? If the former, do industries undergoing exploratory innovation grow faster in terms of market value, output or productivity? Do these industry-wide exploratory episodes result in Schumpeterian dynamics with changes in concentration and higher entry/exit of firms? In the aggregate, what are the dynamics of economy wide exploration-phases? Is there a secular trend in exploration and does this correlate with the much noted slowdown in aggregate productivity growth? At higher frequencies, do we see cyclical movements?

Additionally, we can integrate our measure into state-of-the-art quantitative heterogeneous firm/endogenous growth environments (Aghion, Akcigit, and Howitt, 2014). The latter literature has typically resorted to the use of patent citations as a proxy for patent quality and breadth which provide useful moments to quantifying models featuring "internal versus external" innovation decisions by firms (Akcigit and Kerr, 2018) or "incremental versus disruptive" innovations (Acemoglu, Akcigit, and Celik, 2014). However, patent citation measures have recently been shown to be distorted measures of the value and reach of innovation, be it because of strategic patenting (Abrams, Akcigit, and Popadak, 2013), strategic citations (Lampe, 2012) or measurement error and changes in the way citations are used by patent applicants, both in the cross-section and over time (Kuhn, Younge, and Marco, 2020).

Against this background, our firm-level innovation measures provide an alternative – and arguably more direct – metric to assess how distant a given innovation is relative to the neighbourhood of the knowledge space that has already been visited by a firm; i.e. a measure that preserves a notion of distance without resorting to citations. This, in turn, will enable us to explore quantitative environments that not only take into account the number of 'product lines' a firm currently innovates in, but also where, in the knowledge space, such product lines reside and how, at the micro-level, firms traverse this space over their life-cycle. At the macro-level, such environment further enables us to provide quantitative answers as to whether, for example, secular trends in exploration versus exploitation dynamics at the micro-level can account for a secular productivity growth slowdown.

Appendix 2.A Data

This appendix describes the construction of our data set. Section 2.A.1 discusses the general definition of patent abstracts. Section 2.A.2 and Section 2.A.3 describe our main sources of patent abstracts. Section 2.A.4 describes the procedure we use to webscrape the remaining patents. Section 2.A.5 discusses our text cleaning and pre-processing steps.

2.A.1 Patent Abstracts

We focus on utility patents filed at the United States Patent and Trademark Office (USPTO). More than 90 percent of UPSTO patents belong to the class of utility patents (Bergeaud, Potiron, and Raimbault, 2017). A utility patent provides intellectual property of an invention to its owner. As stated in Title 35 U.S. Code §101:

> *"Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title."*

The general conditions for patentability are novelty (35 U.S. Code §102) and non-obvious subject matter (35 U.S. Code §103). From 1860 to 1995, protection was granted for 17 years. Since 1995 the protection period has been increased to 20 years. According to PCT Rule 8 in the USPTO guidance, an abstract is supposed to be

> *"A summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention."*

2.A.2 Pre-1976 Patent Texts

We obtain the full patent text data for granted patents filed before 1975 from Iaria, Schwarz, and Waldinger (2018). The data set is constructed from digitalised versions of U.S. patents for grant years 1920 to 1979 from the web page of the USPTO. The patent texts were recovered using optical character recognition (OCR) scans and stored in plain text format. Note that the texts obtained from OCR may contain recognition errors introduced during the process of translating from image to text. These are typically caused by imperfections in the original scanned images. As pointed out by Kelly et al. (2018), going backward in time from 1976, the quality of OCR scans generally decreases due to a lower quality typesetting. The final data set is comprised of over 2.5 million patents with a total of more than 7.5 billion words.

Since our analysis focuses on patent abstracts, we extract the abstracts from the full texts where available using regular expressions. In particular, we consider the

following three scenarios. First, if both section titles "abstract of the disclosure" and "background of the invention" can be found in the text, take the abstract as the text between the two titles. Second, in the case that the section title "background of the invention" is not contained in the full text but "abstract of the disclosure" and take the next 150 words as the abstract based on the UPSTO limit of 150 words for patent abstracts. Third, in cases where the abstract is not available, we extract the first 250 words of full text and use them as pseudo-abstracts.

2.A.3 Post-1976 Patent Texts

For patent abstracts of granted patents from 1976 to 2013 we rely on the MongoDB database created by Bergeaud, Potiron, and Raimbault (2017). They obtain the patent texts from USPTO bulk downloads. The total database consists of 4,666,365 utility patent abstracts.

2.A.4 Google Patents

When merging the above pre- and post-1976 data sets we find that they do not contain all patents granted when cross-checking against the list of three million patents from 1963 to 1999 in the NBER legacy data set (Hall, Jaffe, and Trajtenberg, 2001). We webscrape the text of patents that were not included in either of the two above sources from google patents.

2.A.5 Text Cleaning and Pre-Processing

After merging the three sources, we conduct a series of text cleaning and pre-processing steps. We begin by to converting terms into their linguistic roots. In particular, we extract word stems from the patent abstracts using the NLTK Snowball Stemmer. Note that the resulting word stems are not necessarily proper English words. We then use regular expressions to remove numbers and other non-alphabetic characters. Next, we remove occurrences of common stop words defined as terms that with little semantic content such as prepositions and pronouns appearing frequently in all texts.

This is followed by filtering out extremely rare or frequent words. Intuitively, frequent words are used in a majority of patents which in turn renders them uninformative with respect to a specific invention. At the same time, including rare words that are not integral to identifying a technology considerably increases the computational costs when applying our exploration and exploitation measures. For this purpose, we compute the term frequency–inverse document frequency (tf-idf) scores for each remaining keyword in each document. We use a sublinear (logarithmic) transformation to reduce the influence of extremely large or small scores. To reduce the size of the vocabulary, we remove all terms with a tf-idf score lower than

0.1. Finally, we eliminate all patents without any words left in their corpus after the previous removal step. The resulting data sample contains a total number of 277,019 distinct words.

Appendix 2.B Approximate Inference

For a given collection of documents, the inferential problem is to compute the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) \;=\; \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)},$$

where $\boldsymbol{\theta}$, $\boldsymbol{z}$, and $\boldsymbol{w}$ denote the corpus-level sets of the respective document parameters. This posterior distribution is intractable. In the following, we outline the approximate posterior inference procedure. For a more detailed derivation, we refer the reader to Blei, Ng, and Jordan (2003).

The basic idea is to replace the above posterior by a fully factorised variational distribution

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) \;=\; \prod_d q_\theta(\theta_d|\gamma_d) \prod_n q_z(z_{d,n}|\phi_{d,n}),$$

where the variational distribution of the topic proportions $\boldsymbol{\theta}$ is Dirichlet with parameter $\boldsymbol{\gamma}$ and the variational distribution of the topic assignments $\boldsymbol{z}$ is multinomial with parameter $\boldsymbol{\phi}$. This is followed by minimising the Kullback-Leibler (KL) divergence, or relative entropy, between the variational distribution $q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$ and the true posterior $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$. Note that minimising the KL divergence is equivalent to maximising the lower bound on the log likelihood of the observed documents $\log p(\boldsymbol{w}|\alpha, \beta)$ obtained from applying Jensen's inequality. This yields the variational updates

$$\begin{aligned} \phi_{d,n} \;&\propto\; \beta_{w_{d,n}} \exp(\mathbb{E}_q[\log(\theta_d)|\gamma_d]) \\ \gamma_d \;&=\; \alpha + \sum_n \phi_{d,n}. \end{aligned}$$

The variational updates have the following intuitive interpretation. The multinomial update corresponds to using Bayes' Theorem to obtain $p(z_n|w_n) \propto p(w_n|z_n)p(z_n)$. In the update equation, $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution. The update for the Dirichlet parameter is a posterior Dirichlet computed by adding the expected observation counts under the variational distribution $\mathbb{E}_q[z_n|\phi_n]$ to the pseudo-counts α (Blei, Ng, and Jordan, 2003). Using an Expectation Maximisation (EM) algorithm to maximise the variational lower bound yields the approximate empirical Bayes estimates. Specifically, the E-step consists of maximising the lower bound with respect to the variational paramters θ and γ. In the M-step, the bound is maximised with respect to the model parameters α and β. In our application to patent texts, we rely on the online variational Bayes implementation of LDA provided by the Gensim Python library.

Appendix 2.C Additional Firm Figures

Table 2.C.1. 1-year Changes in Sales and Exploration.

	(1) Baseline	(2) +SIC4	(3) +$\Delta_1$ln(PAT)$_t$	(4) +Age
Exploration$_{t-1}$	0.00160***	0.00138***	0.00132***	0.000786**
	(0.000238)	(0.000248)	(0.000248)	(0.000259)
$\Delta_1$ln(PAT)$_t$			0.00837***	0.00911***
			(0.00234)	(0.00233)
age				-0.00209***
				(0.000340)
age2				0.0000152***
				(0.00000340)
R-sq	0.048	0.067	0.067	0.070
N	22,738	22,732	22,732	22,732

	(1) Baseline	(2) +SIC4	(3) +$\Delta_1$ln(PAT)$_t$	(4) +Age
SuccessX$_{t-1}$	0.00273***	0.00244***	0.00226***	0.00216***
	(0.000390)	(0.000386)	(0.000384)	(0.000383)
$\Delta_1$ln(PAT)$_t$			0.0111***	0.0116***
			(0.00331)	(0.00331)
age				-0.00215***
				(0.000344)
age2				0.0000152***
				(0.00000335)
R-sq	0.055	0.079	0.080	0.084
N	19,835	19,826	19,826	19,826

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of the 1-year log change in firms sales $\Delta_1$ln(Sales)$_t$ on the 1-year lag of the general Exploration measure (top) and Successful Exploration (bottom). Year effects in all regressions, SIC4 fixed effects from col(2) onwards. $\Delta_1$ln(PAT)$_t$ is the 1-year change in log patent numbers log(1+PAT).

Table 2.C.2. 5-year Changes in Sales and Average Lagged Exploration.

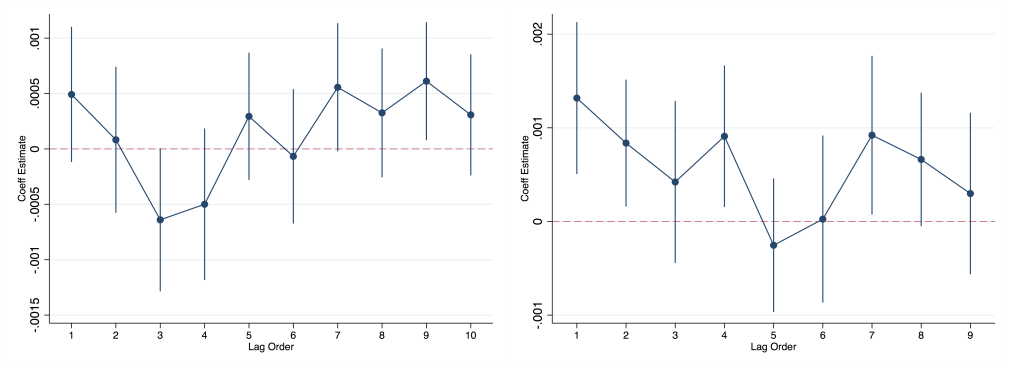
Panel (A)

	(1) 1st-5-years	(2) 10-years	(3) +$\Delta_5$ln(PAT)	(4) all-available
Exploration$_{(t_6-t_{10})}$	0.00414	-0.000277	0.000848	0.00257
	(0.000212)	(0.00229)	(0.00224)	(0.00162)
Exploration$_{(t_{11}-t_{15})}$		0.00667***	0.00471*	0.00441**
		(0.00192)	(0.00198)	(0.00139)
$\Delta_5$ln(PAT)$_t$	0.0363***	0.0357***	0.0368***	0.0378***
	(0.00465)	(0.00466)	(0.00464)	(0.00446)
$\Delta_5$ln(PAT)$_{t-6}$			0.0223*	
			(0.00883)	
$\Delta_5$ln(PAT)$_{t-11}$			0.0189**	
			(0.00640)	
R-sq	0.210	0.213	0.316	0.134
N	10,865	10,865	10,865	20,719

Panel (B)

	(1) 5-years	(2) 10-years	(3) +$\Delta_5$ln(PAT)	(4) all-available
SuccessX$_{(t_6-t_{10})}$	0.00991***	0.00976***	0.00907***	0.00976***
	(0.00228)	(0.00227)	(0.00260)	(0.00227)
SuccessX$_{(t_{11}-t_{15})}$		0.00486*	0.004	0.00486*
		(0.00246)	(0.00267)	(0.00246)
$\Delta_5$ln(PAT)$_t$	0.0421***	0.0421***	0.0423***	0.0421***
	(0.00511)	(0.00512)	(0.00507)	(0.00512)
$\Delta_5$ln(PAT)$_{t-6}$			0.00522	
			(0.0102)	
$\Delta_5$ln(PAT)$_{t-11}$			0.00630	
			(0.00898)	
R-sq	0.229	0.230	0.230	0.230
N	10,140	10,140	10,140	10,140

Notes: Standard errors clustered by firm in parentheses. This table shows the results of regressions of $\Delta_5$ln(Sales)$_t$ on general Exploration and Successful Exploration ('SuccessX'). The exploration measures are included as 5-year averages over the intervals of $(t_6 - t_{10})$ and $(t_{11} - t_{15})$. $\Delta_5$ln(PAT)$_t$ is the 5-year change in log patent numbers log(1+PAT) in period t. The 'All available' column in Panel (A) allows for taking averages in cases where all five 1-year lags are not defined. In Panel (B) this is the same as Column (2) since SuccessX requires continuous data in order to be defined.



(a) Ten years of Lagged Exploration **(b)** Ten years of Lagged Successful Exploration

Figure 2.C.1. One-Year Changes in Sales and Lagged Exploration.

Note: This figure shows the estimates of a regression of the 1-year log change in firm sales on (simultaneous) lags of the general and successful exploration measures. Stand errors clustered by firm and 95% confidence intervals reported.

References

Abrams, David, Ufuk Akcigit, and Jillian A. Popadak. 2013. "Patent Value and Citations: Creative Destruction or Strategic Disruption?" *SSRN Electronic Journal*, DOI: 10.2139/ssrn.2351809. [69]

Acemoglu, Daron, Ufuk Akcigit, and Murat Alp Celik. 2014. "Young, Restless and Creative: Openness to Disruption and Creative Innovations." *SSRN Electronic Journal*, DOI: 10.2139/ssrn.2392109. [69]

Acemoglu, Daron, Ufuk Akcigit, and William R. Kerr. 2016. "Innovation network." *Proceedings of the National Academy of Sciences* 113 (41): 11483–88. DOI: 10.1073/pnas.1613559113. [47]

Aghion, Philippe, Ufuk Akcigit, and Peter Howitt. 2014. "What Do We Learn From Schumpeterian Growth Theory?" *Handbook of Economic Growth* 2: 515–63. DOI: 10.1016/B978-0-444-53540-5.00001-X. [69]

Akcigit, Ufuk, and William R. Kerr. 2018. "Growth through heterogeneous innovations." *Journal of Political Economy* 126 (4): DOI: 10.3386/w16443. [69]

Andrews, Michael J., and Alexander Whalley. 2021. "150 Years of the Geography of Innovation." *Regional Science and Urban Economics*, (December): 103627. DOI: 10.1016/j.regsciurbeco.2020.103627. [57, 59, 64]

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez. 2018. "Text matching to measure patent similarity." *Strategic Management Journal* 39 (1): 62–84. DOI: 10.1002/smj.2699. [39]

Balsmeier, Benjamin, Mohamad Assaf, Tyler Chesebro, Gabe Fierro, Kevin Johnson, Scott Johnson, Guan Cheng Li, Sonja Lück, Doug O'Reagan, Bill Yeh, Guangzheng Zang, and Lee Fleming. 2018. "Machine learning and natural language processing on the patent corpus: Data, tools, and new measures." *Journal of Economics and Management Strategy* 27 (3): 535–53. DOI: 10.1111/jems.12259. [39]

Barron, Alexander T.J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, institutions, and innovation in the debates of the French Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (18): 4607–12. DOI: 10.1073/pnas.1717729115. arXiv: 1710.06867. [37, 39, 44]

Bergeaud, Antonin, Yoann Potiron, and Juste Raimbault. 2017. "Classifying patents based on their semantic content." *PLoS ONE* 12 (4): 1–22. DOI: 10.1371/journal.pone.0176310. arXiv: 1612.08504. [45, 70, 71]

Berkes, Enrico. 2018. "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts." [46, 47]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1. [36, 39, 73]

Bowen, Donald, Laurent Fresard, and Gerard Hoberg. 2021. "Rapidly Evolving Technologies and Startup Exits." *Working paper*, [39]

Bradshaw, R., and C. Schroeder. 2003. "Fifty years of IBM innovation with information storage on magnetic tape." *IBM Journal of Research and Development* 47 (4): 373–83. DOI: 10.1147/rd.474.0373. [49]

Bussy, Adrien, and Friedrich Geiecke. 2020. "A Geometry of Innovation." *SSRN Electronic Journal*, (September 2019): 1–63. DOI: 10.2139/ssrn.3676831. [39]

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. "Reading tea leaves: How humans interpret topic models." *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–96. [41]

Cohen, Wesley M. 2010. *Fifty years of empirical studies of innovative activity and performance*. Vol. 1, 1 C. Elsevier B.V., 129–213. DOI: 10.1016/S0169-7218(10)01004-X. [36]

Cyert, Richard M., and James G. March. 1963. *A Behavioral Theory of the Firm*. Prentice-Hall. [36]

Dennis, Wayne. 1956. "Age and Productivity among Scientists." *Science* 123 (3200): 724–25. DOI: 10.1126/science.123.3200.724. [36]

Ellison, Glenn, and Edward L. Glaeser. 1997. "Geographic concentration in U.S. manufacturing industries: A dartboard approach." *Journal of Political Economy* 105 (5): 889–927. DOI: 10.1086/262098. [63]

Galenson, D. W., and B. A. Weinberg. 2000. "Age and the quality of work: The case of modern American painters." *Journal of Political Economy* 108 (4): 761–77. DOI: 10.1086/316099. [36]

Griliches, Zvi. 1990. "Patent Statistics as Economic Indicators: A Survey." *Journal of Economic Literature* 28 (4): 1661–707. [35]

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." [45, 47, 71]

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production." *Quarterly Journal of Economics*, (June): 927–91. DOI: 10.1093/qje/qjx046. Advance. [45, 70]

Itti, Laurent, and Pierre Baldi. 2009. "Bayesian surprise attracts human attention." *Vision Research* 49 (10): 1295–306. DOI: 10.1016/j.visres.2008.09.007. [36, 37, 39, 41, 42]

Jones, Benjamin, E.J. Reedy, and Bruce A. Weinberg. 2014. "Age and scientific genius." URL: https://www.fatherly.com/wp-content/uploads/2016/03/w19866.pdf. [36]

Kaltenberg, Mary, Adam B Jaffe, and Margie E Lachman. 2021. "Invention and the Life Course: Age Differences in Patenting." *National Bureau of Economic Research Working Paper Series* No. 28769: URL: http://www.nber.org/papers/w28769%7B%5C%%7D0Ahttp://www.nber.org/papers/w28769.pdf. [36, 47, 66]

Kelly, Bryan T., Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2018. "Measuring Technological Innovation over the Long Run." *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3279254. [39, 70]

Kogan, Papanikolaou, Seru, and Stoffman. 2017. "Technological innovation, resource allocation, and growth." *Quarterly Journal of Economics*, (November): 665–712. DOI: 10.1093/qje/qjw040.Advance. [46]

Kuhn, Jeffrey, Kenneth Younge, and Alan Marco. 2020. "Patent citations reexamined." *RAND Journal of Economics* 51 (1): 109–32. DOI: 10.1111/1756-2171.12307. [69]

Lampe, Ryan. 2012. "Strategic Citation." *Review of Economics and Statistics* 94 (1): 320–33. DOI: 10.1162/REST_a_00159. [69]

Lehman, H. C. 1960. "The age decrement in outstanding scientific creativity." *American Psychologist* 15 (2): 128–34. [36]

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2020. "IPUMS National Historical Geographic Information System: Version 15.0." [47]

March, J.G. 1991. "Exploration and exploitation in organizational learning." *Organization Science* 2 (1): 71–87. DOI: 10.1287/orsc.2.1.71. arXiv: z0009. [36]

Moretti, Enrico. 2012. *The new geography of jobs*. New York: Houghton Mifflin Harcourt Publishing Company. [57]

Moretti, Enrico. 2019. "The effect of high-tech clusters on the productivity of top inventors." *NBER Working Paper No. 26270*, [57]

Murdock, Jaimie, Colin Allen, and Simon Dedeo. 2017. "Exploration and exploitation of Victorian science in Darwin's reading notebooks." *Cognition* 159: 117–26. DOI: 10.1016/j.cognition.2016.11.012. [39, 42]

Nicholas, Tom. 2015. “Scale and Innovation During Two U.S. Breakthrough Eras Scale and Innovation During Two U.S. Breakthrough Eras.” [36]

Packalen, Mikko, and Jay Bhattacharya. 2015. “New Ideas in Invention.” *Working Paper*, DOI: 10.3386/w20922. [39]

Syed, Shaheen, and Marco Spruit. 2018. “Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation.” *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017* 2018-Janua: 165–74. DOI: 10.1109/DSAA.2017.61. [45]

Chapter 3

Endogenous Technology Space*

3.1 Introduction

How do firm innovation profiles compare within and across industries? This question plays a central role in the study of innovation and industrial organisation. Long-standing literatures in these areas have documented technological spillover effects and identified industry boundaries. Traditional approaches, however, have mostly relied on pre-defined, exogenous classification systems under the implicit assumption that the relationship between the technological similarity of firms and their closeness in the product market is fixed over time. Furthermore, from a practical perspective, maintaining, improving and updating such classification systems requires large amounts of resources.

In this paper, we use probabilistic machine learning to efficiently construct a new endogenous technology space from patent texts. We then rely on information-theoretic methods to construct measures of technological distance at the firm level – both fixed and time-varying. The fixed distances are comparable to previous approaches relying on patent classes to estimate research and development (R&D) spillover effects. Our new time-varying distances for each year allow us to investigate how the relationship between firm innovation profiles changes both within and across industries.

In doing so, we present three sets of findings. First, we observe that industries are becoming more technologically specialised and segregated over time. The magnitude of this development is masked when using fixed patent classes. Second, we identify the emergence of internet companies in the mid-1990s as a distinct group of firms with roots in traditional information and communication technologies. Third,

* I thank Vasco M. Carvalho, Mirko Draca, and Bill Janeway for helpful comments and suggestions. I thank participants at the Cambridge-UCL Empirical Micro PhD Seminar for useful feedback and suggestions. I gratefully acknowledge the financial support of The Alan Turing Institute under research award No. TU/C/000030.

we determine the unique set of time-varying rivals surrounding a focal firm in the endogenous technology space. We demonstrate the validity of this approach with a case study of the software company Oracle.

Our empirical implementation starts from roughly 1.9 million patent abstracts from 1970 to 2008 matched to approximately 6700 individual firms. We aggregate the abstracts to documents at the firm or firm-year level for the fixed and time-varying firm distances, respectively. We then construct our endogenous technology space and measure firm distances in two steps. First, we apply Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) to the firm corpora. LDA is a hierarchical Bayesian model that allows us to infer the technologies directly from the patent texts. This yields an interpretable, lower-dimensional technology space and allows us to probabilistically represent each firm's position as a mixture of technologies.

In a second step, we compute the distinctiveness between firms using the Jensen-Shannon distance (JSD). The JSD is a proper distance function based on the Jensen-Shannon divergence and as such grounded in information theory. Using a proper metric has major theoretical advantages over previous approaches in the literature mostly relying on the cosine similarity (or uncentred correlation) between vectors of shares to measure firm relatedness. For instance, distances in the endogenous technology space correspond to our usual intuition of spatial distances. It also allows us to apply common dimensionality reduction techniques and clustering algorithms that require a proper distance function such as multidimensional scaling and Hierarchical Agglomerative Clustering (HAC).

Next, we provide a qualitative assessment and baseline validation of our approach. We find that our technology space exhibits a considerable amount of industry structure based on the Standard Industrial Classifications (SIC). We confirm this relationship in a dyadic regression. We then apply HAC to obtain firm technology clusters and illustrate the technological within- and across-industry cluster heterogeneity. In a dyadic logistic regression set-up, we provide evidence that the firm clusters in the endogenous technology space capture some amount of information on the industry classifications.

The qualitative assessment is followed by using our measure to identify general industry trends. Starting from the technology clusters, we identify the emergence and vanishing of technologies. Specifically, we observe the creation of new information and communication technologies (ICT) in the 1990s while the share of firms belonging to clusters describing analogue and mechanical technologies decreases. Next, we estimate and examine the time-varying technological firm distances. In general, we find that the distance matrices illustrate the emergence of rather sharply defined technological industry clusters over time. That is, innovation becomes more segregated with respect to firm SICs. In a dyadic regression setting, we observe a significant decline in distance between firms belonging to the same SIC industry over

time for both series starting from around 1985 – again implying that technology is becoming more industry specific. We also verify these observations using the time-varying industry classifications by Hoberg and Phillips (2016) derived from product descriptions filed with the Securities and Exchange Commission (SEC).

Looking at the industry trends in more detail, we observe three large blocks forming over time relating to three different SIC ranges. The first block is a sharply defined cluster of firms in the chemical industry already clearly identifiable in the 1980s. The second block we observe is a heterogeneous cluster of ICT technologies starting from the 1990s with clearly defined sub-blocks in 2000. Third, we observe the emergence of the cluster of internet companies classified under the non-descript SIC-based "business services" industries. Our findings corroborate the results by Hoberg and Phillips (2016) that many of these firms address rather distinct product markets "using the internet". We show that the firms are in fact highly technologically similar and segregated from other firms while maintaining a relationship to the traditional ICT industries. Our text-based technological distances allow to better identify this development compared to the class-based distances. This is because the patent class system does not accommodate these rapid technological shifts since innovations naturally result in the need to revise existing classifications (Lafond and Kim, 2019).

Lastly, we provide a case study of a company in the "business services" industry, specifically the software company Oracle, to illustrate the depth of the information captured by our measure. We make two main observations. First, we are able to illustrate the emergence of some of Oracle's largest technological competitors. Second, our approach allows us to identify a handful of firms amongst the closest technological rivals that were later acquired by Oracle. This type of information can be used, for example, to investigate synergy effects in mergers and acquisitions.

Related Literature

Our paper connects to the strand of research in innovation and industrial organisation that empirically estimates firm proximity in technology and product market spaces. Methodologically related to our text-based approach, Hoberg and Phillips (2016) rely on annual firm 10-K product descriptions filed with the Securities and Exchange Commission (SEC) to construct an endogenous product market space, time-varying measures of firm similarity and a corresponding set of industry classifications. Their year-by-year set of product similarity measures are then defined as the cosine similarities between the firm-level word frequencies obtained from the product descriptions. They find evidence that firm R&D and advertising are associated with subsequent differentiation from competitors.

The seminal paper by Jaffe (1986) is the first to construct technology spaces from patent data. In particular, Jaffe constructs the technology space from firm-level patent class distributions and defines firm proximity in terms of the cosine similarity between class share vectors. He then uses this technological firm proximity measure to quantify the effects of other firms' R&D spending on the productivity of a focal firm's R&D, that is, spillovers of R&D. His results suggest that a firm's R&D productivity is increased by the R&D of technological neighbours. At the same time, however, he finds that neighbours' R&D lowers the profits and market value of low-R&D-intensity firms.

Bloom, Schankerman, and Van Reenen (2013) formalise this observation of firm performance being simultaneously affected by a positive effect from technology spillovers and a negative business stealing effect from product market rivals. For this purpose, they extend the Jaffe framework by constructing the product market space in a methodologically similar fashion. In particular, to determine a firm's position in the product market space, they calculate the cosine similarity between the firm's sales shares across four-digit SIC industries. They find that the positive effect of technology spillovers quantitatively dominates to the extent that the gross social returns to R&D are at least twice as high as the private returns.

Lucking, Bloom, and Van Reenen (2019) update the results by Bloom, Schankerman, and Van Reenen (2013) using a longer panel data set that includes the beginning of the 21st century. They find that the estimated technology and product market spillovers have been comparable to the earlier results. For the period from 1995 to 2005, they observe larger technology spillovers and smaller negative product market spillover effects which they interpret to reflect the market exuberance for high-R&D firms during the digital technology boom.

Our paper also contributes to a small but growing literature outside of economics that uses probabilistic machine learning models in combination with information-theoretic quantities such as the Jensen-Shannon Distance to analyse social phenomena from text data sources. For example, Jing, DeDeo, and Ahn (2019) apply LDA to obtain topics from an online fanfiction corpus. They then measure topic novelty as the JSD between a fanfiction's topic distribution and the centre of the feature space.

The remainder is organised as follows. Section 3.2 introduces our methodology and provides the baseline validation of our measure. Section 3.3 uses the firm distances to investigate industry trends and technological change. Section 3.4 concludes.

3.2 Methodology

This section introduces our methodology. Section 3.2.1 describes how we span the technology space from patent texts. Section 3.2.2 defines our technological distance measure and discusses the advantage of our methodological framework compared to other possible approach. Section 3.2.3 illustrates how our approach can be used to estimate R&D spillovers. Section 3.2.4 describes our data set. Section 3.2.5 provides a baseline validation of our measure and estimates technology clusters.

3.2.1 Spanning Technology Space from Patent Texts

To span a low-dimensional technology space, we use Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003). LDA is a hierarchical Bayesian model that allows us to infer the technologies directly from the patent texts. Each technology is described in terms of a mixture of technical components, that is, the words found in all patents. Compared to alternative approaches such as calculating the cosine similarities between high-dimensional word vectors, the advantage of LDA is that it facilitates the interpretability of our technology space. In particular, LDA allows us to probabilistically represent each firm as a mixture of technologies corresponding to a point in the low-dimensional technology space. Thus, from observing the firms' positions we can readily characterise their innovation activities.

Intuitively, LDA identifies technologies by finding patterns of co-occurrence of technological components among innovation activities across firms. Specifically, assume that all firms share a set of observed technological components denoted by X constructed from the entire collection of patent texts. A technology k is defined to be a probability distribution β_k over this shared set of technological components X. The innovation activity of a firm i is then described by a mixture θ_i over all technologies K. Formally, a collection of firm patent corpora is generated as follows:

(1) For each firm i:

   a. Draw technology proportions $\theta_i|\alpha \sim \text{Dir}(\alpha)$.

   b. For each technological component $x_{i,n}$:

      i. Draw assignment $z_{i,n}|\theta_d \sim \text{Mult}(\theta_d)$.

      ii. Draw technological component $x_{i,n}|z_{i,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{i,n}})$,

where z_d denotes the technology assignments and α is a K-dimensional Dirichlet parameter. Note that the generative model can be equivalently expressed based on the annual firm patent corpora. We rely on the online variational Bayes algorithm implemented in the Gensim Python library for parameter inference.

We use our method to construct both a fixed and a time-varying version of the firm-to-firm distance matrix. The difference between the two is the definition of

the documents in the LDA inferential procedure. In particular, for the fixed distance matrix, we aggregate the patent texts at the firm level yielding a technology vector θ_i for each firm i in the sample period. To obtain a time-varying distance matrix that allows to capture changes over time, the patent texts are aggregated at the firm-year level instead. In this way, we infer technology vectors $\theta_{i,t}$ for each firm i in year t. Note that this implies that the size of the distance matrix varies over time depending on the number of firms in each given year.

3.2.2 Measuring Distance between Firms

We employ information-theoretic methods to measure the distance between firms in the technology space. Specifically, we compute the distinctiveness between firms based on the Jensen–Shannon divergence (JS). Let p and q be two probability distributions and $m = \frac{1}{2}(p+q)$ their uniform mixture distribution. The Jensen–Shannon divergence is given by

$$\mathrm{JS}(p||q) \;=\; H(m) - \frac{1}{2}\big(H(p) + H(q)\big)$$

where $H()$ is the Shannon entropy (Shannon, 1948). When using a logarithm with base two, it is measured in bits and bounded from above by one for two probability distributions (Lin, 1991).

For two firms i and j with technology distributions θ_i and θ_j, respectively, $\mathrm{JS}(\theta_i||\theta_j)$ measures how much information one sample of patent texts contains on average regarding the distinguishability between the firms. Intuitively, this is computed as all the bits of information in which they each differ from their uniform mixture distribution $\frac{1}{2}(\theta_i + \theta_j)$. Hence, if two firms use the same technologies we are not able to distinguish them and the information is zero. If they differ a substantially, we need a lot of bits to communicate this outcome and describe all their distinctions.

The square-root of the Jensen–Shannon divergence is a proper distance function – we refer to this as the Jensen-Shannon Distance (JSD). That is, it is symmetric, equal to zero for two firms with the same mixture of technologies and it satisfies the triangle inequality (Lin, 1991). Matching the spatial interpretation of a technology space, we define the pairwise distance between two firms i and j as

$$D_{ij} \;:=\; \sqrt{\mathrm{JS}(\theta_i||\theta_j)}.$$

From Distance to Proximity

Note that previous approaches in the literature measure similarity or proximity. Distance, on the other hand, is a measure of dissimilarity. Corresponding to the proximity measure in the definition of technology spillovers by Jaffe, we compute the

proximity between two firms i and j as

$$P_{ij} := 1 - D_{ij}.$$

Defined in this way, P_{ij} has properties equivalent to the proximity measure in Jaffe (1986). Specifically, it yields unity for firms whose text-based technology vectors are identical, that is, where the distance D_{ij} between them is equal to one. In case there is no technological overlap in two firms' innovation activities, it is equal to zero. Thanks to the Jensen-Shannon distance being bounded between zero and one, P_{ij} has the same bounds as Jaffe's measure.

Theoretical Advantages

In addition to its interpretability, the way we construct the lower-dimensional technology space and measure the distance between firms has theoretical advantages over previous approaches in the literature. The most common approach is to calculate the cosine similarity (or uncentred correlation) between two vectors of shares and impose an interpretation of a spatial distance. For example, Jaffe (1986) calculate the cosine similarity between patent class shares to calculate technological proximity between firms. Bloom, Schankerman, and Van Reenen (2013) extend Jaffe's measure to other vectors and calculate, amongst others, the geographical distances between firms. Hoberg and Phillips (2016) calculate the cosine similarity between word frequency vectors. They refer to the word frequency vectors as a spatial location and create industry classifications based on the closest rivals. They compare this to distances on a map.

A straightforward extension in the context of patent texts would be to calculate the cosine similarity between word frequency vectors obtained from the firms' patent portfolios. This approach to measuring the "distance" between firms, however, has theoretical drawbacks due to the fact that the cosine similarity is not a proper metric (Dongen and Enright, 2012). Formally, the cosine similarity does not satisfy the triangle inequality, which demands that the shortest path between two firms is a direct line. Intuitively, the cosine similarity only measures how close two firms are in terms of the angle between the directions of their share vectors as seen from the origin without taking into account their actual distance from each other in space.[1] This has obvious implications for creating industry, technology or geographical classifications based on the closeness between firms.

As pointed out above, the JSD is a proper distance function and as such satisfies the triangle inequality. Hence, distances between firms in the technology space correspond to our usual intuition of distance. Note, however, that the JSD, in contrast

1. To illustrate this, consider the following geographical example: viewed from London as the origin, Paris is closer to Sydney than Berlin in terms of angular distance.

to the cosine similarity, is not independent of the dimensionality of the technology space. In our application, the dimensionality is directly determined by pre-defining the number of technologies when spanning the technology space. Hence, for a smaller number of technologies, we measure the distance between firms in fewer dimensions compared to a larger number of technologies. As noted by Klingenstein, Hitchcock, and De Deo (2014), the JSD behaves sensibly when coarse-graining the technology space from an information-theoretic perspective. In particular, the JSD between firms in a lower-dimensional technology space is smaller or equal to the distance in higher dimensions. If subtle technological distinctions do not contain much information on the distinctiveness between two firms, then their JSD in the lower-dimensional technology space will be close to the high-dimensional case. If, however, a larger number of technologies is particularly useful to distinguish between the two firms, then coarse-graining the technologies to a lower-dimensional space will significantly reduce the distance between them. In our empirical application, we run a number of robustness checks regarding the dimension of the technology space.

3.2.3 Technology Spillovers

Before qualitatively assessing our text-based approach, we briefly outline how our measure of distance in the endogenous technology space carries over to the estimation of technology (or knowledge) spillover effects as commonly found in the growth, productivity, and industrial organisation literatures.

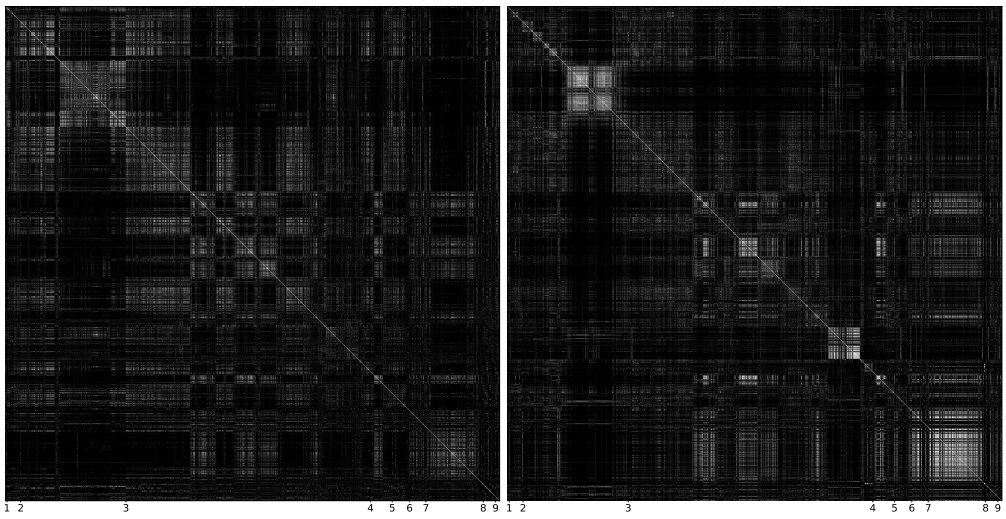
Jaffe (1986) provides the earliest use of patent classes to construct a measure of proximity in the technology space. He then uses his proximity measure to estimate R&D spillovers from neighbouring firms. Building on this framework, our text-based proximity measure can be used to define technology spillovers to a focal firm i as

$$S_i = \sum_{j \neq i} P_{ij} R_j,$$

where R_j is firm j's R&D spending and P_{ij} is our text-based proximity measure between i and j.

3.2.4 Data

We construct our firm-level data set from various sources. We obtain the roughly 1.9 million patent abstracts from 1970 to 2008 matched to 6697 firms from Carvalho, Draca, and Kuhlen (2020). This data set also contains information on the Standard Industrial Classification (SIC) up to the four-digit level. For more details on the text pre-processing steps, we refer the reader to their data appendix. In addition, we obtain information on the United States Patent Classification (USPC) of each patent from Berkes (2018). We combine this data with the classification of USPC patent



(a) Text-Based Distances **(b)** Class-Based Distances

Figure 3.1. Firm Distances.

Notes: The figure shows the technological distance between firms for the period from 1970 to 2008 based on the patent texts (left) and on the pre-defined patent classes (right). Each row indicates the distance of a firm to each of the other firms where lighter shades imply low distance while darker shades represent larger distances between pairs of firms. Firms are sorted by four-digit SIC code with the first digit indicated on the abscissa.

classes into technological categories and sub-categories following Hall, Jaffe, and Trajtenberg (2001). We obtain the mapping for this from Acemoglu, Akcigit, and Kerr (2016).

3.2.5 Qualitative Assessment and Baseline Validation

To begin with, we create a fixed representation of textual firm distances in the technology space. This construction is comparable to previous approaches building on a fixed patent classification. In particular, from the above data set, we compute the text-based firm distance matrix for the 6697 firms from 1970 to 2008 using 50 topics. The documents in the LDA inferential procedure are the patent abstracts aggregated at the firm level. Note that our results do not change significantly for different numbers of topics. For comparison, we also create the distance matrix based on the pre-defined patent classes. For this purpose, we calculate the shares of patents across patent classes aggregated at the 36 subcategory level for each firm. Relying on the sub-categories rather than the 428 classes ensures that the dimensions of the spanned technology spaces are in a similar range. The class-based firm distance is then defined as the JSD between the sub-category share vectors.

In general, we observe a correlation of 0.3 between the text- and class-based pairwise firm distances. Figure 3.1 shows the two matrices with firms sorted by four-

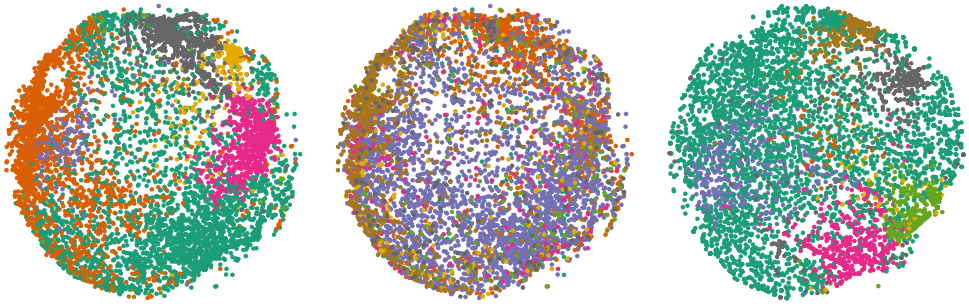
Table 3.1. Effect of Same SIC Industry on Technological Distance.

	(1) Text-Based	(2) Class-Based
Intercept	0.9271***	0.9259***
	(0.00003)	(0.00003)
Same SIC dummy	-0.1221***	-0.2652***
	(0.00001)	(0.00001)
Observations	22,421,556	22,414,860

Notes: Dependent Variable: Firm Distance. This table shows the results from regressing the distance between two firms on a dummy variable indicating whether they belong to the same four-digit SIC industry.

digit SIC industry codes. Lighter shades imply high similarity while darker shades represent dissimilar pairs of firms. The diagonal represents the distance of a firm to itself and thus is white. The most prominent finding is that the both the text-based and class-based measures show a similar structure. In particular, we observe several 'blocks' in the distance matrices. These blocks represent firms that tend to be more similar to firms from the same block compared to all other firms. Given that the firms are sorted by their industry classification, this implies that firms belonging to similar industries work on more similar technologies, which makes intuitive sense. The shade of the blocks, however, differs between the two matrices. Specifically, the class-based graph shows two very light, more sharply defined blocks at the top left and bottom right. These represent the chemical (SIC 2) and business services (SIC 7) industries. The text-based industry blocks show more structure for the large block of manufacturing and ICT firms (SIC 3) in between.

To further illustrate this relationship of the text-based and class-based technology clusters with the industry classifications of firms, we run the following simple dyadic regression. First, we extract the upper triangle of the distance matrix (excluding the diagonal) to obtain pairwise firm distances. We then construct a dummy indicating whether two firms belong to the same four-digit SIC industry. Table 3.1 shows the results. For the text-based clusters in Column (1), we find a clear negative relationship indicating that firms belonging to the same industry are on average around -0.12 closer together in the endogenous technology space. For the patent class-based space, we find that the coefficient on the same SIC dummy decreases to -0.27. Thus, on the surface, it seems as if the information obtained from patent classes overlaps more with the SIC industry classifications. In the following section, however, we show that this relationship is reversed once we account for changes over time.



(a) Text-Based **(b)** Industry Classification **(c)** Class-Based

Figure 3.2. Multidimensional Scaling and hierarchical Clustering.

Notes: The figure shows the results from applying multidimensional scaling and hierarchical clustering. Figure 3.2a shows the two-dimensional MDS representation based on the Jensen-Shannon technology distance matrix with ten clusters indicated by the colours. Figure 3.2b shows the same two-dimensional representation but with colours indicating one-digit SIC industries. Figure 3.2c is based on the Jensen-Shannon distance between patent class and ten clusters.

Hierarchical Technology Clusters

Since the above procedure yields a symmetric firm distance matrix, we can readily apply standard clustering and dimensionality reduction techniques to infer technology clusters. One standard clustering procedure in the statistics and machine learning literature is hierarchical clustering. Hierarchical clustering can be performed either top-down or bottom-up. Since it is more commonly used, we focus on the bottom-up algorithm known as Hierarchical Agglomerative Clustering (HAC) (Manning, Raghavan, and Schütze, 2008). Intuitively, in each iteration, HAC merges the two most similar clusters starting from a proper distance matrix in the first step. Based on the properties of the JSD, we can therefore apply HAC to find technological firm clusters. The distance between the clusters can be computed using different linkage functions, the most common ones include single-link (maximum similarity), complete-link (minimum similarity), and average-link.

We now illustrate using hierarchical clustering to identify firm clusters. We rely on the complete linkage function in the HAC algorithm. To visualise the resulting clusters, we employ multidimensional scaling (MDS). Given a dissimilarity matrix based on a proper distance function, MDS allows us to project the relationship between the firms to a two-dimensional space (Mead, 1992). While there are rules of thumbs to choose the optimal number of clusters, the advantage of using HAC is that it provides a full hierarchy of clusters. This enables us to visualise our results for a small number of clusters first. Specifically, we begin our qualitative assessment based on ten clusters. Figure 3.2 shows the resulting graphs.

Table 3.2. Dyadic Logistic Regression.

	(1) Text-Based	(2) Class-Based
Intercept	-2.5525***	-2.6179***
	(0.001)	(0.001)
Same Cluster Dummy	0.9750***	1.6915***
	(0.003)	(0.002)
Observations	22,421,556	22,414,860

Notes: Dependent variable is same two-digit SIC dummy. This table shows the results from a dyadic logistic regression of a dummy indicating whether two firms belong to the same two-digit SIC industry on a dummy indicating whether they belong to the same text-based or class-based cluster. The probabilities stated in the main text are calculated based on the logit link function. Standard errors in parentheses.

Starting from the left, Figure 3.2a shows the two-dimensional representation for the text-based distance matrix. The graph shows distinct clusters with firms of the same cluster being close to each other. This indicates that our approach is able to identify groups of firms that are similar to each other in terms of their patent texts. Note that the dark green cluster, while spatially close in some places, seems to capture a large group of firms that do not belong to any other cluster and would probably be divided into several smaller sub-clusters for a larger number of clusters. For illustration, Figure 3.2b shows the same two-dimensional text-based distance representation but with colours indicating the ten distinct one-digit SIC industry codes. Comparing this to the ten clusters obtained from HAC, we find that there is some overlap between the industry classification and the text-based firm clusters. Thus, visually our measure of technological distance seems to contain some amount of information on the fixed industry classifications. Lastly, Figure 3.2c shows the results for patent categories and ten clusters. We again find that the clustering algorithm is able to identify clusters that are spatially related. Similar to the text-based clusters, there seems to be one dominant cluster which could be split into smaller clusters.

For this reason, we increase the number of clusters in the following. Specifically, based on computational optimisation methods, we set the number of clusters in the HAC algorithm to 50 for both the text- and class-based. Hence, the number of clusters is in a similar range as the 72 distinct two-digit SICs. To quantitatively assess the overlap between the clusters and industry memberships, we construct the following two dummy variables for a dyadic regression setup: a dummy indicating whether two firms belong to the same two-digit SIC and a dummy indicating whether two firms belong to the same class- or text-based cluster, respectively.

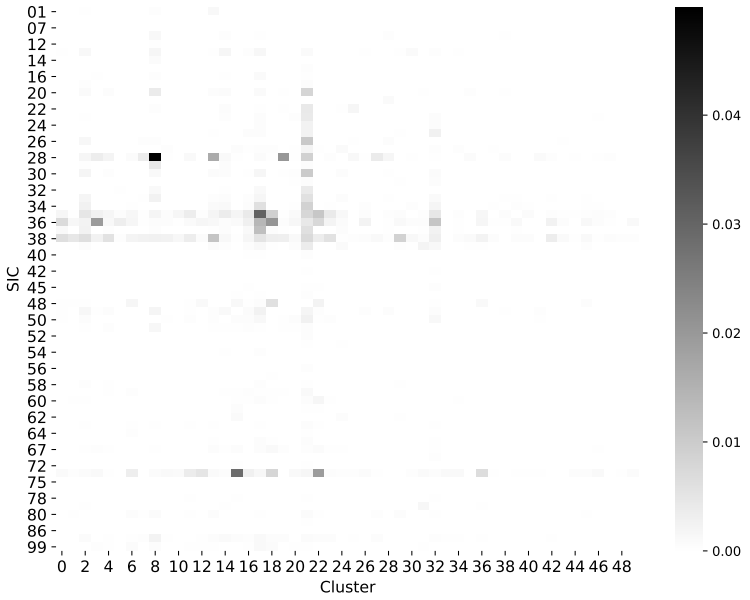


Figure 3.3. Distribution over Industries.

Note: This figure shows the distribution of text-based clusters over two-digit SIC industries. The shade of a cell represents the share of firms that belong to both the particular SIC industry and the cluster indicated on the abscissa.

Table 3.2 shows the results for a logistic regression of the same-industry dummy on the same-cluster dummies. For the text-based clusters in Column (1), we find that the probability of firms in the same class-based cluster also belonging to the same two-digit SIC industry is around 28%. The probability for the text-based cluster membership is a more modest 17%. Hence, while distinct, both patent classes and texts capture some information on the fixed industry classification of a firm.

Next, we investigate the distribution of the text-based clusters over industries. Figure 3.3 shows the frequency of firm cluster memberships across two-digit SICs. This allows us to visually inspect the within- and across-industry heterogeneity with respect to the technological clusters. We make three main observations.

First, we find that ICT firms usually classified under SIC 35, 36, and 38 are spread across clusters. This within-industry heterogeneity across firm technology clusters implies that ICT firms are technologically related to many different fields. That is, using the fixed firm distances from the entire sample period, most clusters contain a certain share of ICT firms.

Second, we observe a highly similar pattern for SIC 73 (“Business Services”). Interestingly, while overall similar, we also see that, for example, cluster 15 seems to capture business services firms that are technologically distinct from traditional ICT firms in the cross section. We will further investigate the relationship between

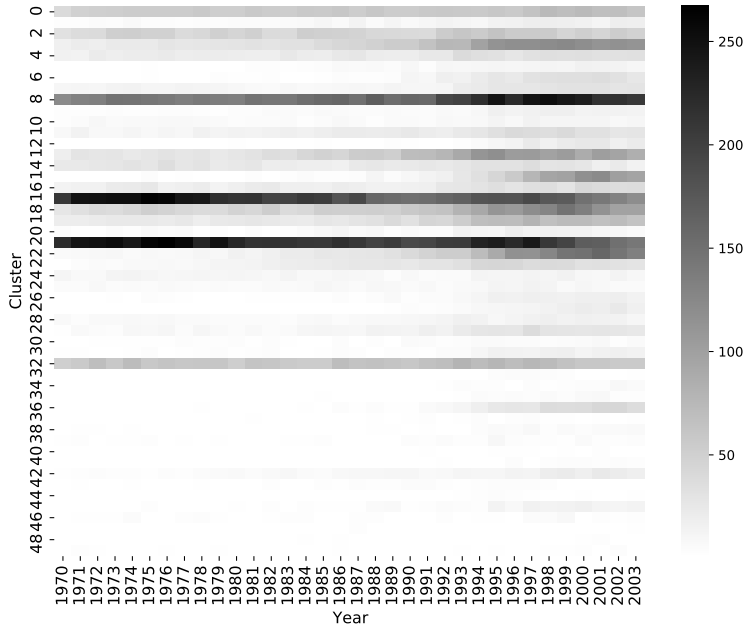


Figure 3.4. Evolution over Time.

Note: This figure illustrates the development of cluster memberships over time. The shade of a cell indicates the number of firms that belong to a specific cluster in a given year.

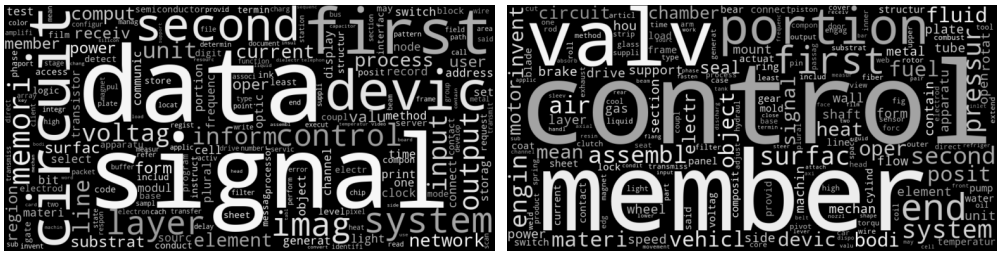
these industries using the time-varying firm distances in the following section where we show that these firms in fact are internet companies that emerged in the mid-1990s.

Third, the graph also allows us to examine the across-industry heterogeneity of the technology clusters. Most prominently, cluster 21 is visibly spread across industries falling into the full range of SIC 20 to 39. This range of SIC codes represents the division of "Manufacturing" firms covering a broad range of technologies in areas such as food, textile, furniture and ICT. Hence, this cluster captures the share of manufacturing firms with common technological elements across these fields.

The results presented in this section so far provide prima facie evidence of both text- and class-based firm distances in the endogenous technology space capturing some information related to the fixed industry classifications. The next section takes a dynamic perspective and uses our measure to investigate technological and industry changes.

3.3 Capturing Technological Change and Industry Trends

This section applies our methodology to investigate industry trends and technological change. Section 3.3.1 examines changes in technology cluster memberships over time. Section 3.3.2 estimates the time-varying firm distance matrices and describes



(a) Emerging Clusters. **(b)** Vanishing Clusters.

Figure 3.5. Word Clouds for Emerging and Vanishing Text-Based Clusters.

Notes: The figure shows the word clouds for emerging clusters (3, 11, 13, 15, 22, and 36) and vanishing (17 and 21) text-based clusters.

the development of industry patterns over time. Section 3.3.3 presents evidence on the emergence of internet companies in the mid-1990s. Section 3.3.4 provides a case study of the software company Oracle.

3.3.1 Development of Technology Clusters

We start our analysis by examining changes in the technology clusters constructed in the previous section. This allows us to get a first insight into major technological changes captured by our text-based measure. Keeping clusters constant over time facilitates the interpretation compared to reestimating the clusters for each year.

Figure 3.4 illustrates the changes in the cluster membership distributions over time. To allow for a dynamic interpretation, we compute the distribution of firms over clusters for a given year. Note that the cells now represent the absolute number of firms to facilitate the comparison between different years. We generally observe a clear pattern of the emergence of a few clusters starting in the early to mid-1990s. At the same time, we observe another group of clusters becoming smaller over time. Cluster 21 plays a prominent role again as part of the latter group, indicating that the absolute number of manufacturing firms has decreased over time.

To get a better understanding of these trends, we compute the word clouds for the emerging clusters (3, 11, 13, 15, 22, and 36) and vanishing clusters (17 and 21). Figure 3.5 shows the two word clouds. We see that the emerging clusters are comprised of firms working in digital technologies most prominently characterised by words such as “signal” and “data”. The vanishing clusters on the other hand consist of firms innovating in analogue and mechanical technologies described by words such as “valve”, “pressure”, and “vehicle”. Hence, the development of memberships over time indicate a decline in traditional manufacturing firms and a technological shift during the digital technology boom in the 1990s.

3.3.2 Industry Trends

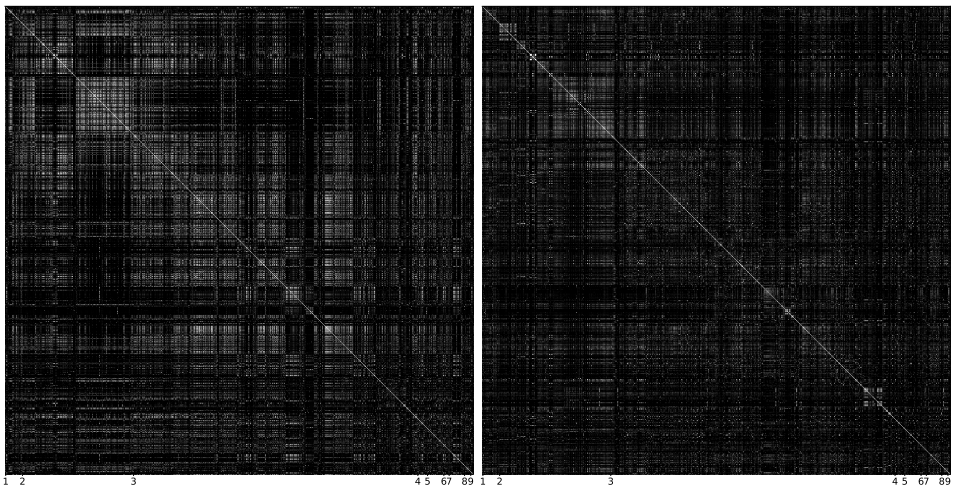
A major advantage of our text-based approach is the ability to provide a time-varying distance matrix. To identify the dynamic changes in firm distances, we re-estimate the endogenous technology space as follows. First, we aggregate a firm's patent text by the application year. We then span the technology space by applying LDA to the firm-year text corpus. From this, we obtain a technology distribution describing each firm's patenting activities in a given year. We then compute our time-varying firm distances based on the firm-year technology distributions for each year. For comparison, we also calculate annual patent-class shares aggregated at the subcategory level and use the JSD to compute the class-based technological distance measure.

Figure 3.6 shows both the text- and class-based distance matrices for the years 1980, 1990 and 2000. Firms are again sorted by four-digit SIC code with the first digit indicated on the abscissa. Both distance matrices illustrate the emergence of distinct technological clusters over time.

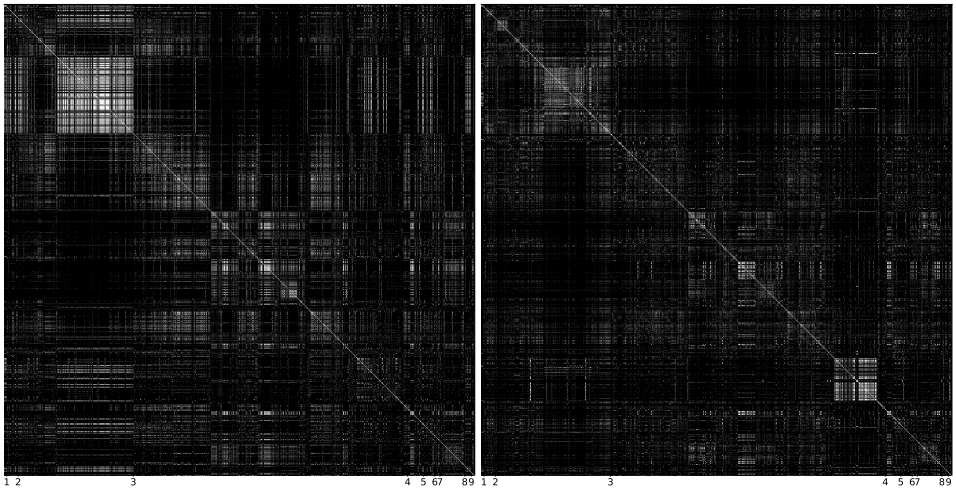
Specifically, in 1980, both the text- and class-based distance matrix look rather uniform. The text-based distance matrix shows the beginnings of larger blocks, particularly in the SIC 2 range. In 1990, these larger industry blocks are clearly visible. The overall appearance of the class-based distances shows a resemblance to the text-based matrix. Lastly, in 2000, the text-based distance matrix exhibits distinct industry blocks along the diagonal as does the class-based graph, albeit not as sharply defined. In summary, the general trend we observe in both graphs implies that firms belonging to the same industry are becoming more similar in terms of their innovation efforts over time. This process of industry specialisation and technological segregation is identified earlier using the patent texts compared to the pre-defined patent classes.

Next, we investigate these industry trends and their relation to each other in more detail. Based on the block structure in the distance matrices, we make three main observations relating to three different SIC ranges. First, the sharply defined large block at the top left falling in the range of one-digit SIC 2 mainly represents firms belonging to SIC 28 ("Chemical & Allied Products"). This cluster of firms in the chemical industry is already clearly identifiable in the 1980s, especially in the text-based graph. We observe that this block is particularly "self-contained" and thus segregated in the sense that these firms are only technologically similar to other firms from SIC 28 and rarely related to firms from other industries.

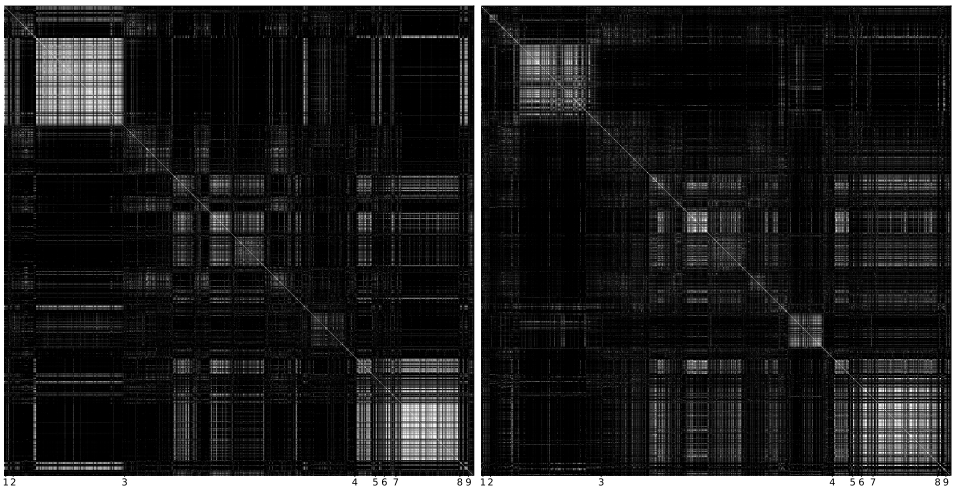
Second, the large structure in the middle of the graph for the year 2000 is comprised of firms in the one-digit SIC 3 class. The two largest groups belong to the two-digit SIC codes 35 ("Industrial Machinery & Equipment"), 36 ("Electronic & Other Electric Equipment"), and 38 ("Instruments & Related Products") thus representing ICT firms. The graph illustrates the emergence of ICT technologies with



(a) Text-Based Distances 1980 **(b)** Class-Based Distances 1980



(c) Text-Based Distances 1990 **(d)** Class-Based Distances 1990



(e) Text-Based Distances 2000 **(f)** Class-Based Distances 2000

Figure 3.6. Firm Distances.

Notes: The figure shows the text-based and class-based technological distance between firms for the years 1980, 1990 and 2000. Each row indicates the distance of a firm to each of the other firms where lighter shades imply high similarity while darker shades represent dissimilar pairs of firms. Firms are sorted by four-digit SIC code with the first digit indicated on the abscissa.

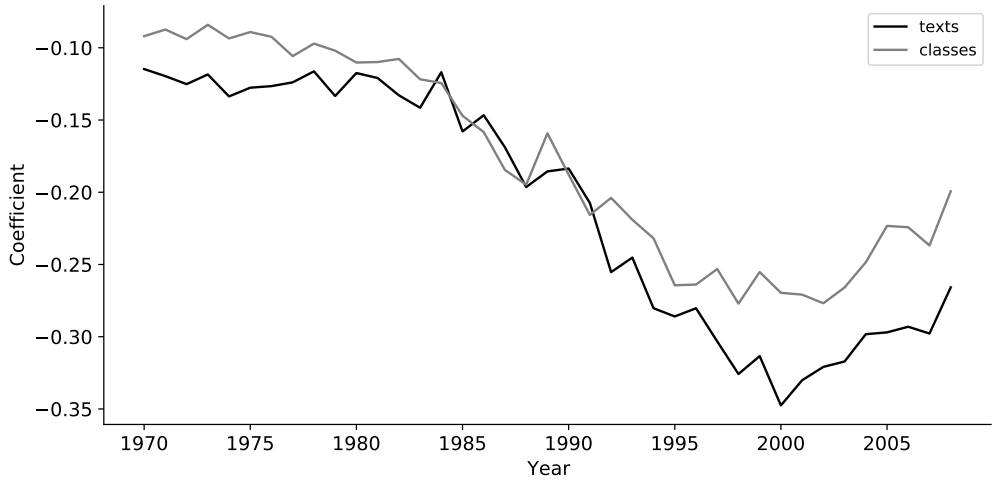


Figure 3.7. Effect of Same SIC Industry on Technological Distance over Time.

Note: This figure shows the coefficient from regressing the class-based (grey) and text-based (black) technological distances on a dummy variable indicating if two firms belong to the same industry four-digit SIC from 1970 to 2008.

the block formations beginning already in the 1980s. This development leads to clearly defined blocks in 2000. In contrast to the block in the chemical industry, the boundaries are less sharp and there is more heterogeneity in firm distances. Firms in the main industries SIC 35 and 36, however, are strongly technologically related to each other. We note that the class-based matrix shows a slightly more clearly defined block at the end of the SIC class 3 range of firms belonging to SIC 38 ("Instruments & Related Products") compared to the text-based measures.

Third, we observe the emergence of the cluster of firms belonging to the two-digit SIC 73 ("Business Services") and more specifically SIC 737 ("Computer and data processing services") at the three-digit level. The block visibly starts forming in the 1990s both in size and in technological relatedness. Before analysing this development in greater detail in Section 3.3.3, however, we first investigate the general relationship of our technological distance measure to industry classifications.

Relationship to Industry Classifications

Similar to the analysis in Section 3.2.5, we run a dyadic regression of both text- and class-based firm distances on a dummy variable indicating whether two firms belong to the same four-digit SIC industry for each year. Figure 3.7 displays the resulting time series of the dummy variable coefficients. Table 3.A.1 and Table 3.A.2 in Appendix 3.A contain the corresponding regression results for the years 1970, 1980, 1990, and 2000. The graph shows a significant decline in distance between firms belonging to the same four-digit SIC industry over time for both series starting

Figure 3.8. Correlation Between Technological and Text-Based Product Similarity.

Note: This figure shows the correlation coefficient between both the class- and text-based technology similarity measures (defined as one minus technological distance) and the text-based product similarity score by Hoberg and Phillips (2016) for each year from 1988 to 2008.

from around 1985. This again implies that technology is becoming more industry specific. The trend is more pronounced for the text-based series which reaches its minimum in 2000 with a coefficient of around -0.35. That is, in 2000, two firms belonging to the same industry are on average 0.35 closer in the technology space. Note that this coincides with the height of the Dot-Com Boom. For comparison, the class-based coefficient is around 0.25. After 2000, however, we see a slight reversal on the magnitude of the coefficient for both series. Note that we also observe that the intercept in the dyadic regression increases slightly from 0.88 to 0.91 implying an average increase in the distance between industries and thus higher segregation.

To check the robustness of these results, we have additionally run the regressions for one-, two- and three-digit SIC codes. We found that the shape does not change substantially. The magnitude, however, decreases for shorter SIC codes and thus higher levels of aggregation. This makes intuitive sense since we average over an increasing number of firms that do not necessarily work on similar technologies.

Next, we investigate this trend with alternative, time-varying industry classifications. In particular, one drawback of static zero-one membership industry classifications such as SIC is that they do not capture changes over time. This represents a potential source of bias. Hoberg and Phillips (2016) propose time-varying industry classifications derived from product descriptions filed with the Securities and Exchange Commission (SEC). Specifically, they rely on the cosine similarity between word frequency vectors obtained from annual 10-Ks filed with the SEC to identify clusters of firms. We obtain the cosine distances between firms (identified by gvkey)

from Hoberg and Phillips (2016) and merge them with our dyadic technology distance data. Similar to the case of Jaffe's technology spillover measure, the text-based product similarity scores between two firms measure proximity rather than distance. Thus, to ensure comparability, we rely on our definition of firm proximity described in Section 3.2.2, that is, firm similarity equals one minus technological distance in the following analysis.

Figure 3.8 shows the correlation of both the class- and text-based technology similarity measures with the text-based product similarity score by Hoberg and Phillips (2016) for each year from 1988 to 2009. We see a sharp increase in the correlation coefficients for the both class- and text-based measures from around 0.14 at the beginning of the 1990s to 0.36 in 1995. Note that the correlation of the text-based technological similarity measure with firm product similarity is significantly larger than the patent class-based measure for all years except the first.

In general, these results again support the observation that the innovation profiles of firms operating in the same product market have become more similar over time resulting in more technologically specialised and segregated industries. Additionally, they suggest that, while the fixed class-based distances contain more information on the fixed industry classifications, the text-based distances are stronger correlated with the SIC industries when taking into account changes over time. This relationship is even stronger between our time-varying technological distances and the text-based industry distances. Next, we focus on a specific sub-area of the product market space.

3.3.3 Emergence of Internet Companies

As described above, we observe the emergence of the large cluster of firms belonging to "Computer and Data Processing Services" (SIC 737) at the three-digit level falling under "Business Services" (SIC 73). This block grows substantially both in size and technological relatedness in the mid-1990s. Comparing the text- to the class-based graph in Figure 3.6, we note that the SIC 737 block in the text-based distance matrix is more distinct from the surrounding SICs and lighter in shade for firms belonging to this specific SIC implying shorter within industry distances. Additionally, we note that firms in SIC 737 are technologically close to firms in the ICT block described above.

To get a general understanding of the relationship to other industries, Figure 3.9 visualises the development of the share of firms across one-digit SIC codes that SIC 737 firms are close to. For this, we set the upper limit of closeness in the endogenous technology space to 0.2. The graph demonstrates that up until the mid-1990s, SIC 737 firms were mainly technologically close to SIC 3 firms. At the three-digit level, we find the largest groups to be SIC 357 ("Computer and Office Equipment"), 366

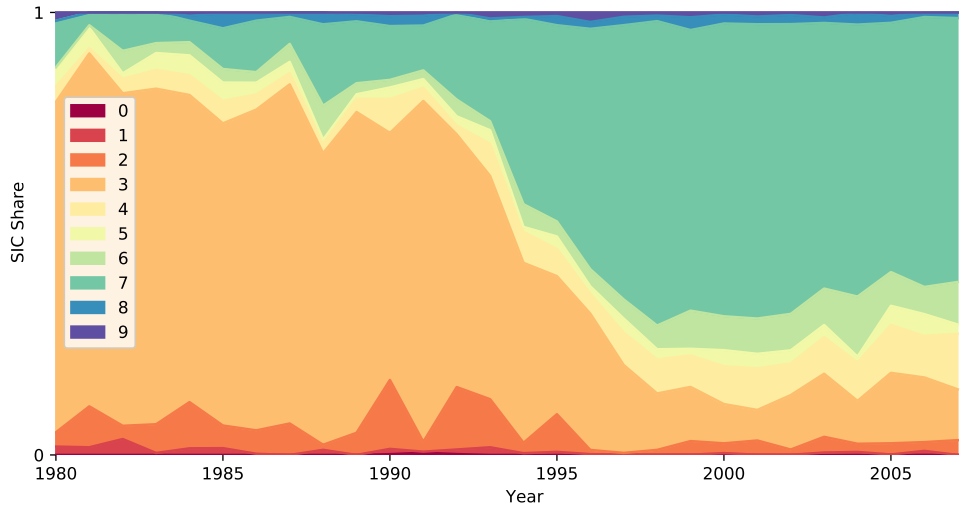


Figure 3.9. Evolution of SIC Shares of Firms Close to Firms in "Business Services".

Note: This figure illustrates the evolution of the shares of one-digit SIC across firms close to firms in SIC 737 from 1970 to 2008. The the upper distance threshold is set to 0.2.

("Communications Equipment") and 382 ("Measuring and Controlling Devices") representing ICT firms. We then observe a major shift in the industry shares with SIC 7 (or SIC 737 at the three-digit level) representing the largest group. Thus, this implies that as the number of firms classified in the Business Services category increased, these firms tended to be technologically most similar to other firms from the same industry and to some degree to firms from traditional ICT classes.

Our findings correspond to the observations by Hoberg and Phillips (2016) in the context of time-varying text-based industry classifications. Specifically, they note that in 1997, a large number of firms were grouped into the non-descript business services industry SIC 737. They point out that accurately specifying industry composition in these types of industries in which firms offer highly differentiated products is particularly difficult. Their results suggest that many of the firms classified in SIC 737 address rather distinct product markets (including entertainment, medical services, information transmission, software, corporate data management and computing solutions) together with firms from SIC 357, 366 and 382. Their overall finding is that the firms address these markets "using the internet" while also competing with rivals that have a more traditional brick-and-mortar presence.

Our results corroborate these findings by providing evidence of the technological relatedness between these firms. Specifically, we find that 1997 represents one of the years during the major shift from technological closeness to SIC 3 firms to SIC 7 firms. Thus, together with the observations by Hoberg and Phillips (2016), we conclude that the rapid growth of SIC 737 represents the new generation of internet firms

in the 1990s that are technologically homogeneous but address different product markets. Put differently, SIC 737 firms are close in the endogenous technology space but more dispersed in the product market space.

Our text-based technological distances allow to better identify this development compared to the class-based distances. This is because the patent class system does not accommodate these rapid technological shifts since innovations result in the need to revise existing classification almost by definition. In particular, Lafond and Kim (2019) find that 40% of the patents granted in 1976 patents currently belong to a different class compared to when they were first issued. After that, the reclassification rate exhibits a sharp decline, reaching about 10% for the 1990's. It is almost zero thereafter. This illustrates that the classification system takes a long time after a radical invention to change. In contrast, our text-based approach incorporates new technological trends in real-time.

Our analysis so far has focused on describing aggregate industry trends. Next, we provide a case study of a firm belonging to SIC 737 to demonstrate the validity of our approach by identifying technological neighbours.

3.3.4 Case Study: Oracle

Our text-based distance measure has allowed us to identify firms classified under SIC 737 as the emerging block of technologically related internet companies. In this section, we use our measure to examine the local development of one of these firms' closest neighbours in the endogenous technology space. In particular, we focus on the enterprise software company Oracle. Oracle is a provider of cloud services and database management applications. Table 3.3 and Table 3.4 show the top 25 neighbours of Oracle as measured by the text-based firm distances for the periods from 1997 to 2001, and 2002 to 2006, respectively. Table 3.A.3 and Table 3.A.4 in Appendix 3.A show the equivalent rankings for the class-based measure.

We make two main observations. First, the tables illustrate the emergence of some of Oracle's largest current competitors. Specifically, we see that Amazon moved close to Oracle in the technology space towards the end of the 1990s. This reflects Amazon's development of their distributed systems and cloud services, which was first used internally and later represented the main revenue stream for the company. Note that Amazon is not identified by the class-based distances to be a close neighbour of Oracle. Other notable competitors of Oracle include Alphabet, Microsoft and SAP which are highly ranked in both the text- and class-based tables starting from 2003.

Second, we observe a handful of firms that were later acquired by Oracle. For instance, we find Siebel Systems, a software company that developed customer relationship management applications and was acquired by Oracle in 2005. We also

Table 3.3. Oracle's Closest Text-Based Competitors.

1997	1998	1999	2000	2001
Novell	Novell	Accenture	Accenture	Mcafee
Adobe	Ca	Novell	Novell	Siebel Systems
Electronic Data Systems	Electronic Data Systems	Ca	Stamps.Com	I2 Technologies
Netscape Communications	Entrust	Amazon.Com	I2 Technologies	Bea Systems
Entrust	Aspect Communications	Entrust	Great Elm Capital Group	Time Warner
Symantec	Netscape Communications	I2 Technologies	Intertrust Technologies	Altaba
I2 Technologies	Concerto Software	Microstrategy	Altaba	Goldman Sachs Group
Intervoice	I2 Technologies	Electronic Data Systems	Ebay	Ca
Sybase	Amazon.Com	Great Elm Capital Group	Entrust	Microstrategy
Aspect Communications	Sap Se	Altaba	Mynd	Great Elm Capital Group
Ca	Liberate Technologies	Genesys Telecomm Labs	Mcafee	Bmc Software
Borland Software	American Management Systems	Symantec	Amazon.Com	Telecommunication Sys
Bmc Software	Nuance Communications	Citrix Systems	Imagex	Palmsource
Amazon.Com	First Data	Dassault Systems Sa	Bmc Software	Vignette
Citrix Systems	Palm	Nuance Communications	Citrix Systems	Akamai Technologies
Sterling Software	Time Warner	American Management Systems	Napster	Overture Services
Time Warner	Lycos	Time Warner	Microstrategy	Amazon.Com
Catalina Marketing	T-Netix	Mcafee	At&T	Sap Se
Bea Systems	Accenture	Sybase	Telecommunication Sys	United Parcel Service
Netspeak	Secure Computing	Siebel Systems	Nuance Communications	Borland Software
Sabre Holdings -Cl A	Verisign	E.Piphany	Alphabet	Citrix Systems
Infoseek	Open Market	Compuware	United Online	Leap Wireless Intl
Lernout & Hauspie Spch Pd	Sybase	Solectron	Ca	Navteq
Merrill Lynch & Co	Blucora	Intertrust Technologies	Fair Isaac	Symantec

Notes: The table shows Oracle's top 25 closest competitors ranked by the Jensen Shannon Distance between text-based technology shares for the years 1997 to 2001.

Table 3.4. Oracle's Closest Text-Based Competitors.

2002	2003	2004	2005	2006
Accenture	Bea Systems	Alphabet	Altaba	Alphabet
Bea Systems	Alphabet	Altaba	Alphabet	Altaba
Siebel Systems	Altaba	Ca	Symantec	Bea Systems
Sap Se	Novell	First Data	Bea Systems	Novell
Novell	Time Warner	Palmsource	Accenture	Accenture
Time Warner	Ca	Ebay	Mcafee	Avaya
Adobe	Accenture	Accenture	Commvault Systems	Mcafee
Genuity	Realnetworks	Novell	Morgan Stanley	Ebay
I2 Technologies	Siebel Systems	Morgan Stanley	Ebay	Red Hat
Ca	Electronic Data Systems	Citrix Systems	Trend Micro	Akamai Technologies
Akamai Technologies	Amazon.Com	I2 Technologies	Citrix Systems	Trend Micro
Great Elm Capital Group	Ebay	Sybase	Federal National Mortga Assn	Morgan Stanley
United Parcel Service	Mcafee	Mcafee	Sap Se	Navteq
Bmc Software	Great Elm Capital Group	Akamai Technologies	Akamai Technologies	Fair Isaac
Altaba	Overture Services	Packeteer	3Com	F5 Networks
Symantec	Metro One Telecomm	Siebel Systems	Intuit	Realnetworks
Enterasys Networks	Vignette	Navteq	Goldman Sachs Group	Nice
Federal National Mortga Assn	Goldman Sachs Group	Trend Micro	Palmsource	Sybase
Capital One Financial	West	Intuit	Red Hat	Microsoft
Intertrust Technologies	Intervoice	Bea Systems	Realnetworks	Sap Se
Opentv	Bmc Software	Bank Of America	Siebel Systems	Capital One Financial
West	Borland Software	West	F5 Networks	West
First Data	Nuance Communications	Sap Se	Novell	Napster
F5 Networks	I2 Technologies	Cognos	Callwave	Ca

Notes: The table shows Oracle's top 25 closest competitors ranked by the Jensen Shannon Distance between text-based technology shares for the years 2002 to 2006.

find BEA Systems, which developed enterprise infrastructure software and was acquired in 2008, to be one of the closest technological neighbours of Oracle. Similarly, Novell is technologically particularly close to Oracle throughout the late 1990s and early 2000s. In 2010, Novell was jointly acquired by a group of companies which Oracle was part of. Note that Novell only occurs twice in the class-based ranking.

In the following, we focus on the development of the relationship between Oracle, BEA Systems and Novell to illustrate the depth of the information captured by our text-based distance measure. Our discussion is based on Janeway (2018), which we refer the reader to for a more detailed analysis of the M&A activities between the three and other adjacent firms.

To begin with, in 1993, the telecommunications company AT&T sold its Unix Systems Laboratory including the distributed transaction processing monitor Tuxedo to Novell. Then, in 1996, BEA acquired all of the commercial and intellectual property rights to Tuxedo as well as the supporting technical resources from Novell. Tuxedo was the crucial technology for BEA's penetration of the enterprise market. BEA subsequently went public in 1997, which is the first year shown in Table 3.3. We see that Novell is the top technological rival of Oracle and BEA is in the top 20. In 2001, 2002 and 2003, we see that BEA is on rank four, two and one of Oracle's closest technological rivals, respectively. The year 2001 marks the peak of BEA's rapid early growth as well as the burst of the Dot-Com Bubble. As described by Janeway (2018), while affected by the general retrenchment in technology markets following the bubble, BEA continued both generating cash flow and investing in new technologies. As a result, BEA continued being in the top 25 closest neighbours of Oracle. In 2006, the final column of our table, BEA is the third closest firm to Oracle in our endogenous technology space. This development reflects the growing establishment of BEA's market leadership in the third wave of distributed computing and finally lead to the acquisition by Oracle in 2008 for more than five times its annual revenue at the time. As mentioned above, Novell, the fourth closest neighbour in 2006, was partly acquired by Oracle two years later in 2010.

3.4 Conclusion

In this paper, we use probabilistic machine learning techniques and information-theoretic quantities to measure firm distances in a new endogenous technology space constructed from patent texts. Using our time-varying firm distances, we provide evidence that industries are becoming more specialised and segregated. We also identify the emergence of internet companies as a technologically distinct cluster of firms with roots in the traditional ICT industry. Finally, we demonstrate the validity of our approach by means of the Oracle case study.

The methodology developed in this paper has a wide range of applications in the innovation and industrial organisation literature that have traditionally relied on patent classes to identify technological neighbours. Similar to the use of text-based product market closeness in the study by Hoberg and Phillips (2010), this type of information can be used, for example, to investigate whether firms exploit synergy effects through technology complementaries in mergers and acquisitions. Furthermore, our measure can also be applied in the context of analysing 'killer acquisitions' as introduced by Cunningham, Ederer, and Ma (2021). This type of acquisition aims at discontinuing a target's innovation projects to terminate nascent sources of threat to the incumbent firm's prospective profits.

In the future, we plan on using our new measures in combination with exogenous shocks such as changes in economic policy to examine their impact on the composition of firm clusters in the endogenous technology space. For example, we have started investigating the question of how increases in defense spending create new areas in the technology space. This builds on the framework by Carvalho and Draca (2017) who analyse the propagation of U.S. military spending shocks along supply chains. The underlying intuition is based on the notion of demand-led innovation. Specifically, the government creates a significant portion of final demand. One important aspect of government-led innovation is defense-spending-led innovation where the government creates a space or market for new and specialised technologies. These types of applications benefit from time-varying firm distances in an endogenous technology space that allow to account for rapid technological changes.

Appendix 3.A Additional Figures

Table 3.A.1. Annual Dyadic Text-Based Regression.

	(1) 1970	(2) 1980	(3) 1990	(4) 2000
const	0.889***	0.898***	0.916***	0.914***
	(0.000)	(0.000)	(0.000)	(0.000)
Same SIC dummy	-0.114***	-0.117***	-0.183***	-0.352***
	(0.004)	(0.003)	(0.002)	(0.001)
Observations	413,595	563,391	715,806	2,001,000

Notes: Dependent Variable: Firm Distance. This table shows the results from regressing the text-based distance between two firms on a dummy variable indicating whether they belong to the same four-digit SIC industry for the year 1970, 1980, 1990 and 2000.

Table 3.A.2. Annual Dyadic Class-Based Regression.

	(1) 1970	(2) 1980	(3) 1990	(4) 2000
const	0.919***	0.930***	0.942***	0.928***
	(0.000)	(0.000)	(0.000)	(0.000)
Same SIC dummy	-0.092***	-0.110***	-0.188***	-0.270***
	(0.003)	(0.002)	(0.002)	(0.001)
Observations	405,450	563,391	707,455	1,965,153

Notes: Dependent Variable: Firm Distance. This table shows the results from regressing the class-based distance between two firms on a dummy variable indicating whether they belong to the same four-digit SIC industry for the year 1970, 1980, 1990 and 2000.

Table 3.A.3. Oracle's Closest Class-Based Competitors.

1997	1998	1999	2000	2001
Sybase	Vitria Technology	Ca	Bea Systems	Vignette
Infoseek	Lycos	Siebel Systems	Bmc Software	Thermwood
Xcerra	E.Piphany	Ask Jeeves	Opsware	Blucora
Xcellenet	Faro Technologies	Neomedia Technologies	Ita Holdings	Quest Software
Sagent Technology	Crossworlds Software	Critical Path	Sap Se	Virage
Sand Technology	Blucora	Cec Entertainment	Stratasys	American Management Systems
Amazon.Com	United Online	Sagent Technology	Bank Of America	Netapp
Scientific Learning	Tricord Systems	Starbase	Lycos	Equifax
Network Computing Devices	At Home	Webmethods	Agile Software	Overture Services
Portal Software	Determine	Fair Isaac	Ask Jeeves	Siebel Systems
Bea Systems	Sybase	Multex.Com	Sears Roebuck & Co	Red Hat
Initio	Siebel Systems	Merrill Lynch & Co	Amdocs	Informatica
Compuware	Vignette	Art Technology Group	Westrock Co	Pharsight
Novell	Serena Software	Prism Technologies Group	Cnet Networks	Saba Software
Standard Register Co	Bmc Software	Auspex Systems	Verity	Novell
Mercury Interactive	Scientific Learning	Inktomi	Reynolds & Reynolds	Altaba
Altaba	Helix Technology	Sap Se	Ca	Amazon.Com
Ca	Electronic Data Systems	Booking Holdings	United Online	Unisys
Unicom	Inktomi	Netapp	Mynd	Bmc Software
I2 Technologies	Perot Systems	Mynd	Netapp	Borland Software
Creative Technology	Exchange Applications	United Online	Amazon.Com	Sap Se
Keystone International	Swk Holdings	Bea Systems	Novell	Ca
Ncr	Interlake	Novell	I2 Technologies	Bea Systems
Arvin Industries	Servicemaster Co	American Management Systems	Unisys	Microsoft

Notes: The table shows Oracle's top 25 closest competitors ranked by the cosine distance between patent class shares for the years 1997 to 2001.

Table 3.A.4. Oracle's Closest Class-Based Competitors.

2002	2003	2004	2005	2006
Sybase	Bmc Software	Alphabet	Verisign	Sap Se
Iona Technologies	Salesforce.Com	Siebel Systems	Checkfree	Symantec
Napster	Digi International	Intellisync	Iron Mountain	Alphabet
Extended Systems	Cnet Networks	Safenet Holding-Redh	Sap Se	Bea Systems
Palmsource	Omnicare	Scientific Learning	Unisys	Altaba
Bank Of America	Altaba	Netezza	Bea Systems	Sybase
Icad	Arcsight	Aspen Technology	Progress Software	Shutterfly
Altaba	Overland Storage	Omniture	Sybase	Amazon.Com
Ebay	Hudson Technologies	Altaba	Goldman Sachs Group	Novell
Stamps.Com	Informatica	Cognos	Alphabet	Microsoft
Ariba	Dun & Bradstreet	Federal Home Loan Mortg	Microsoft	Ebay
Verity	Weight Watchers Intl	Sap Se	Accenture	Ca
Intellisync	Business Objects Sa	Sybase	Novell	Intuit
Kroger Co	Agile Software	Unisys	Ebay	Red Hat
Sears Roebuck & Co	Thomson Reuters	Intuit	Palmsource	Citrix Systems
Compuware	Aol	Goldman Sachs Group	Ca	Morgan Stanley
Ore Holdings	Quest Software	Bmc Software	Symantec	Mcafee
Sap Se	Vignette	I2 Technologies	Mcafee	Cadence Design Systems
Accenture	Genworth Financial	Sourcefire	Garmin	Nuance Communications
Amazon.Com	Siebel Systems	Trimble	Penson Worldwide	Health Grades
Bea Systems	Sap Se	Netapp	Nasdaq	Drugstore.Com
Netapp	Red Hat	Ca	Bgc Partners	Loudeye
Borland Software	Alphabet	Bea Systems	Schwab (Charles)	Bank Of New York Mellon
Siebel Systems	Sonic Solutions	Ebay	Image Sensing Systems	Western Union Co

Notes: The table shows Oracle's top 25 closest competitors ranked by the cosine distance between patent class shares for the years 2002 to 2006.

References

Acemoglu, Daron, Ufuk Akcigit, and William R. Kerr. 2016. "Innovation network." *Proceedings of the National Academy of Sciences* 113 (41): 11483–88. DOI: 10.1073/pnas.1613559113. [89]

Berkes, Enrico. 2018. "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts." [88]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1. [82, 85]

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying Technology Spillovers and Product Market Rivalry." *Econometrica* 81 (4): 1347–93. DOI: 10.3982/ecta9466. [84, 87]

Carvalho, Vasco M., Mirko Draca, and Nikolas Kuhlen. 2020. "Exploration and Exploitation in US Corporate Science." [88]

Carvalho, Vasco M, and Mirko Draca. 2017. "Cascading Innovation." *Working Paper*, [106]

Cunningham, Colleen, Florian Ederer, and Song Ma. 2021. "Killer acquisitions." *Journal of Political Economy* 129 (3): 649–702. DOI: 10.1086/712506. [106]

Dongen, Stijn van, and Anton J. Enright. 2012. "Metric distances derived from cosine similarity and Pearson and Spearman correlations." URL: http://arxiv.org/abs/1208.3145. [87]

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." [89]

Hoberg, Gerard, and Gordon Phillips. 2010. "Product market synergies and competition in mergers and acquisitions: A text-based analysis." *Review of Financial Studies* 23 (10): 3773–811. DOI: 10.1093/rfs/hhq053. [106]

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124 (5): 1423–65. DOI: 10.1086/688176. [83, 87, 99–101]

Jaffe, Adam B. 1986. "Technological Opportunity and Spillovers of R&D : Evidence from Firms ' Patents , Profits , and Market Value." *American Economic Review* 76 (5): 984–1001. [84, 86–88]

Janeway, William H. 2018. *Doing Capitalism in the Innovation Economy*. Cambridge University Press. DOI: 10.1017/9781108558440. [105]

Jing, Elise, Simon DeDeo, and Yong-Yeol Ahn. 2019. "Sameness Attracts, Novelty Disturbs, but Outliers Flourish in Fanfiction Online." arXiv: 1904.07741. URL: http://arxiv.org/abs/1904.07741. [84]

Klingenstein, Sara, Tim Hitchcock, and Simon De Deo. 2014. "The civilizing process in London's Old Bailey." *Proceedings of the National Academy of Sciences of the United States of America* 111 (26): 9419–24. DOI: 10.1073/pnas.1405984111. [88]

Lafond, François, and Daniel Kim. 2019. "Long-run dynamics of the U.S. patent classification system." *Journal of Evolutionary Economics* 29 (2): 631–64. DOI: 10.1007/s00191-018-0603-3. arXiv: 1703.02104. [83, 102]

Lin, Jianhua. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37 (1): 145–51. DOI: 10.1109/18.61115. [86]

Lucking, Brian, Nicholas Bloom, and John Van Reenen. 2019. "Have R&D Spillovers Declined in the 21st Century?" *Fiscal Studies* 40 (4): 561–90. DOI: 10.1111/1475-5890.12195. [84]

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Modern Information Retrieval*. Cambridge, England: Cambridge University Press. [91]

Mead, A. 1992. “Review of the Development of Multidimensional Scaling Methods.” *Journal of the Royal Statistical Society Series D* 41 (1): 27–39. [91]

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (4): 623–56. DOI: 10.1002/j.1538-7305.1948.tb00917.x. [86]