UAB
Universitat Autònoma de Barcelona

Three Essays on Innovation: A Text-Analysis Approach

JOSEPH ALEXANDER EMMENS

UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Three Essays on Innovation: A Text-Analysis Approach

JOSEPH ALEXANDER EMMENS

SUPERVISOR:
HANNES MUELLER

A dissertation submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
on the International Doctorate in Economic Analysis (IDEA),
Department d'Economia i Història Econòmica.

May 15, 2025

For my parents, and both my brothers, without you, I would never have started.
For Lucía, without you, I would never have finished.

Acknowledgements

This Ph.D. thesis is the product of countless hours of hard work, supported by many wonderful academics, friends, family members, and generally incredibly supportive people. I want to acknowledge them here.

First, I owe a great debt of thanks to my supervisor, Hannes Mueller. Your drive to bring out the best in me and my work has fostered a love of academia and a pride in my research that will last far beyond the Ph.D. I would also like to dedicate a special acknowledgement to Christian Fons-Rosen. Your deep passion for good science and your willingness to walk and talk about research provided the example and space I needed to write this thesis. I would like to give a special thank you to Vicente Bermejo: without your guidance, friendship, and mentorship, I never would have started this journey. Any success I may have encountered, or that lies ahead, is in great part due to the guidance of these three.

I was fortunate enough to write this thesis within our fantastic department with many doors on which to knock in times of joy or doubt. Thank you to Inés, David, Pau, Joan, Ada, Fernando, Hanna, and Jordi for your academic guidance and personal support. A very special thank you goes to the IDEA and IAE staff, particularly Àngels, Merce and Angela, for their endless patience and support. I of course thank my fantastic co-authors Dennis, Tomasso, Felix and Stefano. Working with you is a privilege that I hope to continue for many years.

I am indebted to all the IDEA students with whom I have shared this time, but above all, to the group of us who made it through together—Luis, Manuel, Sergi, and Jacob. You are all the proof I will ever need that science is best done in proximity to other great scientists. Thank you for all the hours spent poring over questions, proofs, data, writing and for sharing both the moments of crisis and success. I would also like to mention Gabi, my mentor in guiding me into life as an academic. Finally, to Nico—hopefully, I hope to have passed some of Gabi's wisdom on.

My parents have always been the source of curiosity, creativity and patience from which I have learnt the skills needed to write this thesis. I thank both my broth-

ers, Tom and Reuben, for their presence, encouragement, and love throughout the process. I also thank Jo, Steph, Bety, Kelo, Edu, and my wonderful extended family, both British and Spanish, which continues to grow and with whom I can continue to share this remarkable journey.

Finally, I want to thank Lucía. Your hard work, love, and strength shine through this Ph.D. I will be forever grateful. This Ph.D. has opened exciting doors for us, and I am proud to continue walking through them with you.

Thesis Summary

Innovation plays a central role in tackling modern economic challenges, from stagnant firm productivity growth to climate degradation and worsening health outcomes. Understanding how and why inventors innovate is therefore of vital public policy importance.

This thesis examines innovation through patent texts, leveraging recent advances in text analysis and machine learning. These methods are transforming daily life, popular culture, and science–including the study of innovation. By integrating these techniques with rigorous economic theory, this thesis improves our ability to measure the knowledge held by inventors, teams, and firms and to understand how they produce the innovations needed to tackle tomorrow's challenges.

In Chapter 1, *Teams and Text: Modelling Collaboration Through Patent Documents*, I introduce a novel methodology that integrates inventor teams and their patent texts into a unified framework for studying collaborative innovation. I develop a Bayesian model of Natural Language Processing that captures the scientific division of labour within teams. By combining high-dimensional patent data with a statistical model of teamwork, the method developed allows me to infer each team member's contribution to a patent's knowledge content. I use this to study collaboration dynamics over the life cycle of an inventor's career.

Building on this framework, Chapter 2, *Catalyst or Constraint: The Dual Role of Prior Innovation for Breakthroughs*, examines how prior innovations shape a team's ability to push the innovation frontier forward. I once again apply the model from Chapter 1 to patent text data. In this case mapping inventors, teams, and research fields into a structure known as the knowledge space. I combine this with data on premature team member deaths to provide a quasi-random shock to the research potential of the team. Through a continuous treatment model, I identify how team innovations change as they pivot to more or less advanced research areas. This framework offers a flexible and tractable approach to studying the creation of new research fields, an area largely overlooked in the literature due to a lack of suitable models and data.

In Chapter 3, *From Shares to Machines: How Common Ownership Drives Automation*, we examine three increasingly important economic phenomena: the rise of common ownership in public firms, monopsony power, and the shift toward automated production processes. This chapter is co-authored with Dennis C. Hutschenreiter, Felix Noth, Stefano Manfredonia, and Tommaso Santini. We propose a theory that greater overlap in the stockholders of local labour market competitors drives automation-related innovation. We measure automation using a classification derived from the text of each patent produced at a firm. To estimate a causal effect, we exploit exogenous increases in common ownership due to institutional investor mergers, which provide a quasi-experimental setting. Our findings confirm that when common ownership among local competitors increases, firms expand automation and reduce employment.

Contents

List of Figures

List of Tables

Chapter 1

Teams and Text: Modelling Collaboration Through Patent Documents

Abstract

This paper models the division of scientific labour within inventor teams and shows how this division can be inferred from patent texts using a Bayesian model of Natural Language Processing. I find that as inventors gain experience, they collaborate on more patents per year, but contribute relatively less to each patent compared to junior co-inventors. Increased concentration in contribution shares generally reduces patent value, reflecting a quantity-quality trade-off made by senior inventors. This trade-off allows seniors to accumulate more patents, though, on average, each one has a lower value. However, when junior inventors are of high quality, this negative effect disappears entirely. Importantly, the quality of the senior does not make up for concentrating the contribution on low quality juniors.

I gratefully acknowledge the support of the Spanish Agencia Estatal de Investigación (MCIN/ AEI /10.13039/501100011033) through grant PID2020-114251GB-I00.

1.1 Introduction

Teams drive scientific progress, yet how they organise the production of knowledge remains largely a black box. As innovation stems from the recombination of existing ideas (Weitzman, 1998), understanding how inventors combine to produce knowledge within teams is crucial. Consider the case of Milton Friedman and Anna J. Schwartz's collaboration on their seminal work *A Monetary History of the United States, 1963*. Without additional information about the authors or their work, it would be difficult to infer that Friedman was a theorist while Schwartz was an empiricist. Yet in fact, Friedman's theory required Schwartz's national accounts data analysis to empirically prove his propositions. This illustrates both the challenge and importance of understanding the division of labour within scientific teams.

This paper studies whether inventors organise themselves within teams following a scientific division of labour and whether this division can be inferred from patent texts. Having shown that it can, I ask whether teams in which members contribute equally produce more valuable innovations. I use the following definition of contribution: the share of knowledge components contained in an innovation contributed by a given inventor. For example, if a drone uses mostly electronics engineering, a little knowledge on metals, and some on physics, we might ask which team member provided each part? Importantly, did one inventor alone contribute the electronics, while their co-inventors contributed the relatively smaller parts on metals and physics? Or did all team members contribute equal shares?[1]

To infer these contribution shares, I model collaboration through the lens of the Author-Topic Model (Rosen-Zvi et al., 2012). This model represents a significant contribution to our ability to disentangle the contributions of individual inventors to team knowledge production. I apply this model to a dataset of 1 million USPTO patents, selected to represent a sample of well-connected inventor teams. The approach allows me to present a set of novel descriptive statistics on inventor contributions. In addition, I propose a novel measure for the concentration of team contributions. This measure captures whether a team divides scientific

[1]It is important to note that it does not reflect the effort contributed to many facets of collaboration: running analysis, writing the patent, applying for grants and other administrative tasks. In this sense, the paper is different to Xu, Wu, and Evans (2013), who use a definition of contribution to general scientific work load. They do not, however, observe their measure for patent texts, and instead use scientific papers as their setting.

labour equally, or whether a few team members dominate. I start by describing contributions at an aggregate level. I then move on to present inventor and team level panel regression models, accounting for unobserved heterogeneity across teams, inventors, and time. While this approach does not establish strict causal identification, it provides robust evidence on the patterns linking team organisation to innovation outcomes.

Teams divide contributions differently across technology classes, and I demonstrate that teamwork is getting "flatter" over time. Moving to the team and inventor level, I demonstrate that as inventors gain experience, they collaborate on more patents per year. In contrast, I find that as inventors become more experienced, they contribute less to each individual patent. This points to a diluting effect of gaining experience: senior inventors collaborate on more patents, but contribute less to each one. If the seniors are contributing less, does the concentration of scientific labour onto a few team members shape patent outcomes? Xu, Wu, and Evans (2013) argue that flat teams produce more disruptive academic science. I extend their result by showing that teams in which each member contributes equally produce more valuable patents.

However, I rationalise this result with the fact that seniors and juniors contribute differently to team patents. Concentration appears, in the majority of cases, from juniors contributing more when collaborating with seniors. This alludes to a potential quantity-quality trade-off for senior inventors. As inventors grow in experience they collaborate more frequently, but dilute their contribution to each individual patent. Therefore, if they collaborate with lower quality inventors, they increase the quantity to quality ratio. As such, the quality of the juniors drives patent outcomes. I demonstrate this by showing that concentration in contribution shares correlates with lower patent values. However, this negative effect disappears entirely when the junior inventor is of high quality. In addition, I show that the quality of the senior does not make up for concentrating the contribution on low quality juniors. These results suggest that for seniors looking to increase their patent output, sourcing quality juniors who will take up the contribution share is key. There may be valid reasons, however, for which a senior cannot collaborate with high quality juniors, and these results should be taken as a first step in understanding this complex dynamic.

This paper's contribution lies in developing a novel model that allows for the scientific division of labour within inventor teams, allowing me to truly open the black box around teamwork. By offering a framework to systematically assess

how inventors combine knowledge, I provide new insights into the structure of innovative collaboration. This framework allows me to extend the literature on how collaboration structure within teams determines outcomes, but also presents a powerful yet parsimonious framework to model the knowledge production function.

Related Literature The notion that innovation occurs as a result of the recombination of existing knowledge is well established within economics (Schumpeter, 1939; Weitzman, 1998; Fleming, 2001). The definition of contribution used in this paper relies on innovation being modular, such that each knowledge component can be described individually. An existing literature argues that innovation is modular to varying degrees (von Hippel, 1990; Brusoni and Prencipe, 2001; Ethiraj and Levinthal, 2004; Vakili and Kaplan, 2021). I frame the definition of contribution within this literature, which allows the model to infer which inventor contributed each knowledge component.

Given the rising importance of teamwork, developing empirical methods to decode how teams combine individual profiles is key to understanding their production process (Ahmadpoor and Jones, 2019). I contribute specifically to the empirical literature which looks to disentangle individual contributions to team projects (Bonhomme, 2022; Mindruta et al., 2024). There is a small but important literature using highly specific case studies in which individual inputs are observed (Kahane, Longley, and Simmons, 2013; Devereux, 2018; Weidmann and Deming, 2021). The method proposed in this paper extends this literature, which typically focuses on contributions to quality, by disentangling the knowledge contribution of each member. The most direct method of achieving this currently is to use the contribution statements published at select journals (Sauermann and Haeussler, 2017). Xu, Wu, and Evans (2013) develop a highly flexible method of measuring hierarchy within teams, trained on contribution data, and show that flat teams produce more disruptive innovations. I develop this literature by making use of a topic model to infer who contributed each section. Bandiera et al. (2020) was a breakthrough paper within economics and the use of topic modelling. They demonstrate that worker types can indeed be inferred from text. They identify manager types from the text within a manager's calendar activities and implement a similar model of text analysis used in this chapter.

This paper explains a set of within team dynamics that contribute to the literature on the division of labour within scientific teams. Within this literature,

recent progress has been made in identifying team leaders, as part of the division of scientific labour (Haeussler and Sauermann, 2020; Wu, Esposito, and Evans, 2024). Given the changing knowledge production function, the manner in which teams distribute contributions matters (Agrawal, Goldfarb, and Teodoridis, 2016). As a result, there is a growing and important literature on the role of specialists and generalists within teams (Anderson, 2012; Graves and Kuehn, 2021). Jones (2009) argues that as knowledge accumulates, this incentivises individuals to specialise and hold deeper knowledge on fewer knowledge components. Melero and Palomeras (2015), however, defend the role of generalists as inventors who can contribute to various knowledge components within the innovation. The framework presented here enables us to better understand which components each inventor contributes to, and therefore track the division of labour among inventor teams.

Finally, I frame the conclusions of this paper on how the division of labour drives patent outcomes within the literature on the effect of experience and ageing on innovation. Gingras et al. (2008) showed that the number of scientific papers an inventor collaborates on increases before plateauing, as they increase in age. However, their average ordinal position in the list of authors increases.[2] This suggests that a similar trend to the diluting effect of senior inventors also exists within academic science. Kaltenberg, Jaffe, and Lachman (2023) look at the effect of age on innovation through U.S. patents. They find a hump-shaped relationship between the number of patents collaborated on in a year and inventor age, with a peak in the late 30s. The results presented here are complementary, given that I examine the role of experience alone and not age.

Paper Outline The rest of the paper is structured as follows: Section 2 outlines the statistical model of teamwork; Section 3 describes the data and sampling method used; Section 4 describes the inference method; Section 5 provides a validation test; Section 6 presents the descriptive results; Section 7 presents the full empirical results, and Section 8 concludes.

1.2 Model of Teamwork

In order to disentangle individual contributions to team projects, I present the Author-Topic Model as a novel model of collaboration. The first version of this

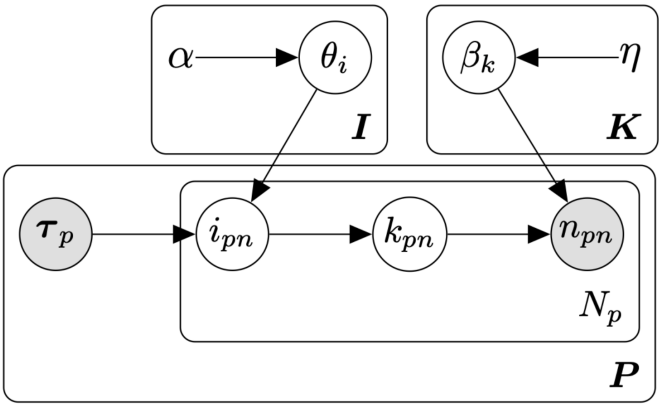
[2]They show that this result holds for normalising by team size, which, given trends of increasing team sizes, may have biased this result.

model was presented in Rosen-Zvi et al. (2012) and I follow the notation laid out in Mortensen (2017). I update their notation to the context of inventors (authors), who use knowledge classes (topics) to produce patents (texts). I refer to both seminal papers for further reference.[3]

A patent $p \in \boldsymbol{P}$ describes an innovation. There exists a set of inventors $\boldsymbol{I}$ who together form teams to produce these patents. Each patent is produced by a team $\boldsymbol{\tau}_p \in \boldsymbol{T}$ of m inventors. For each patent p there is an associated vector $\boldsymbol{n}_p$ of N_p words, where each word n_{pn} is chosen from a vocabulary of size V. There exist $\boldsymbol{K}$ knowledge classes, which represent areas of scientific expertise. The choice of each word n_{pn} is governed by the probability of using each word when describing each knowledge class. The idea being that if you are describing an automation innovation you are more likely to use the words *car, wheel* and *drive* than you are *hospital* or *medicine*.

FIGURE 1.1

INVENTOR-KNOWLEDGE CLASS MODEL



Notes: Plate notation for the baseline Bayesian Hierarchical model behind the Author-Topic Model. This diagram is adapted from Mortensen, 2017.

A collection of patents therefore includes each vector of words and the corresponding team. This is defined formally as: $\{(\boldsymbol{n}_p, \boldsymbol{\tau}_p) \mid p \in \boldsymbol{P}\}$. A set of patents $\boldsymbol{P}$ is produced with the following generative process, assuming a uniform prior over contribution shares. This model is represented in the following plate diagram in Figure (1.1).

- For each inventor $i \in \boldsymbol{I}$ draw $\theta_i \sim \text{Dir}(\alpha)$.

[3]In addition to following the notation laid out in Mortensen, 2017, they also develop the *Gensim* application of the Author-Topic Model. The quantitative model and code for this project were built directly on top of the scripts provided through this python package.

- For each knowledge class $k \in \boldsymbol{K}$ draw $\beta_k \sim \text{Dir}(\eta)$.
- For each document $p \in \boldsymbol{P}$:
  - Given the team $\boldsymbol{\tau}_p$ of patent p
  - For each word in the patent $n \in \{1, \ldots, N_p\}$.
    - Select an inventor for the current word $i_{pn} \in \boldsymbol{\tau}_p \sim \frac{1}{m}$.
    - Conditioned on $i_{pn} = i$, select a knowledge class $k_{pn} \sim \theta_{ik}$.
    - Conditioned on $k_{pn} = k$, select a word $n_{pn} \sim \beta_{kn}$.

Importantly for the objective of this paper, each word in every patent is attached to an inventor, knowledge class pair. This allows the model to infer many important but latent parameters. These include each inventor's knowledge distribution θ_i and the contribution share of inventor i to patent p. The posterior distribution given the observed data and Dirichlet priors is given by

$$P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \alpha, \eta, \boldsymbol{n}, \boldsymbol{T}) = \frac{P(\boldsymbol{n}|\boldsymbol{k}, \boldsymbol{\beta}) P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta}) P(\boldsymbol{i}|\boldsymbol{T}) P(\boldsymbol{\beta}|\eta) P(\boldsymbol{\Theta}|\alpha)}{P(\boldsymbol{n}|\alpha, \eta, \boldsymbol{T})} \tag{1.1}$$

This is the probability of observing the data, the proposed mapping from words to inventor, knowledge class pairs and latent parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. As is typical in Bayesian analysis this posterior is intractable. This is because we have no estimate for the marginal probability of the observed data in the denominator. Therefore topic models use an inference method to back out an approximation. I use a method of Variational Bayes.[4] Define $q(\cdot)$ as an approximation to the posterior

$$q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \lambda, \gamma, \phi) = q(\boldsymbol{\Theta}|\boldsymbol{\gamma}) q(\boldsymbol{\beta}|\lambda) q(\boldsymbol{k}, \boldsymbol{i}|\phi) \tag{1.2}$$

$$\approx P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \alpha, \eta, \boldsymbol{n}, \boldsymbol{T}) \tag{1.3}$$

The definition of the variational approximation introduces an essential feature in order to infer non-uniform contribution shares. The variational approximation introduces three variational parameters, λ γ and ϕ. The first two λ and γ govern the distribution of inventors across knowledge classes, and knowledge classes across words respectively. The key feature is that equation 1.2 now models the choice of knowledge classes and inventors as dependent random variables

[4] Gibbs Sampling is an alternative and popular model, which can give good results, however on large sample sizes can perform very slowly.

where $P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta})P(\boldsymbol{i}|\boldsymbol{T}) \approx q(\boldsymbol{k}, \boldsymbol{i}|\phi)$. This is known in the literature as a blocking estimator. This means that the probability of choosing inventor $i \in \boldsymbol{\tau}_p$ is a function of the knowledge held by inventor i relative to their collaborators, and the knowledge contained in the patent p. In other words, the choice of the inventor and knowledge class are now dependent.

If a patent already includes a lot of words discussing medicine, and if one of the inventors has a larger weight in this knowledge class than others in the team, then they are more likely to be chosen to contribute again. This allows for non-uniform contribution weights $\omega_{ip} \neq \omega_{jp} \; \forall \; i, j \in \tau$ and for the knowledge profile of individual inventors to be over (under) represented in the patent knowledge distribution.

Define the following parametrisation of $q(\cdot)$

$$\begin{aligned} q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\lambda, \gamma, \phi) &= q(\boldsymbol{\Theta}|\boldsymbol{\gamma})q(\boldsymbol{\beta}|\lambda)q(\boldsymbol{k}, \boldsymbol{i}|\phi) \\ &= \prod_i q(\theta_i|\gamma_i) \prod_k q(\beta_k|\lambda_k) \prod_{p,n} q(i_{pn}, k_{pn}|\phi_{ik}) \\ &= \prod_i Dir(\theta_i|\gamma_i) \prod_k Dir(\beta_k|\lambda_k) \prod_{p,n} q(i_{pn}, k_{pn}|\phi_{ik}) \end{aligned}$$

This is the product of the probability of observing I inventor to knowledge class distributions, K knowledge class to word distributions and a set of inventor and knowledge class pairs for each word of every patent. By changing the underlying assumption of how inventors and knowledge classes are drawn to more closely match reality, the plate diagram of parameter dependence changes. Figure (1.2) presents the final model.

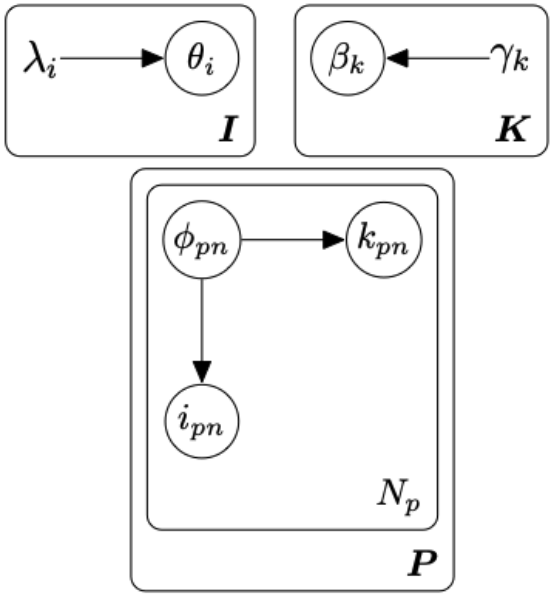
For patent p, the matrix $\phi_{p,v,i,k}$ gives the discrete joint probability of choosing each inventor i and knowledge class k pair for a given word $n = v \in V$. Formally, the probability of inventor i choosing knowledge class k and word v for patent p is given by

$$\phi_{p,v,i,k} = \begin{cases} \phi_{p,v,i,k} & i \in \boldsymbol{\tau}_p \\ 0, & \text{otherwise} \end{cases}$$

The full probability distribution is stored during the estimation as a four dimensional matrix $\phi_{p,v,i,k}$[5]. Where $\sum_i \sum_k \phi_{p,v,i,k} = 1$.

[5]In reality the Gensim package uses the exchangeability of the model to develop an online algorithm to reduce the memory requirements of this matrix, I refer you again to Mortensen, 2017 for further details on this great package.

FIGURE 1.2
INVENTOR-KNOWLEDGE CLASS MODEL: BLOCKED



Notes: Plate notation for Bayesian Hierarchical model in a blocked model, given the assumption that the draw of inventor and knowledge class are dependent, thus allowing for non-uniform contribution shares.This diagram is adapted from Mortensen, 2017.

During inference, the model iterates over each word in every patent, updating the estimates for each of the parameters and the mapping from words to inventor, knowledge class pairs. The method is a derivation of Expectation Maximisation and solves for the following condition using Jensen's inequality.[6]

The right hand side is a lower bound on the marginal probability of the observed data. Also known in the literature as the Evidence Lower Bound (ELBO). A set of functional assumptions allows one to solve the right hand side by defining each of the expected values. The goal is then to maximise this right hand side to approximate the log-likelihood of the observed data as closely as possible. This is implemented in the *Gensim* model using coordinate ascent, which maximises a multivariate function by iterating over each variable and optimising in that direction, holding all others constant until convergence.

$$
\begin{aligned}
log\ p(\boldsymbol{n}|\alpha,\eta,\boldsymbol{T}) \geq \\
\mathbb{E}_q\Big[log\Big(P(\boldsymbol{k},\boldsymbol{i},\boldsymbol{\beta},\boldsymbol{\Theta},\boldsymbol{n}|\alpha,\eta,\boldsymbol{T})\Big)\Big] - \mathbb{E}_q\Big[log\Big(q(\boldsymbol{k},\boldsymbol{i},\boldsymbol{\beta},\boldsymbol{\Theta}|\lambda,\gamma,\phi)\Big)\Big] \\
= \mathcal{L}(\lambda,\gamma,\phi)
\end{aligned}
$$

[6]For a full derivation I refer the reader to the original paper on Latent Dirichlet Allocation by Blei, Ng, and Jordan, 2003.

On convergence, I back out the θ_i given γ_i and β_k given λ_k.[7] The contribution of this paper is to go one step further and use the converged parameter $\phi_{p,v,i,k}$ to back out a contribution share for each team member. To do so, I sum across the relevant dimensions of $\phi_{p,v,i,k}$ as

$$\phi_{p,v,i,k} = \frac{\exp\{\mathbb{E}_q[\log \theta_{ik}] + \mathbb{E}_q[\log \beta_{kv}]\}}{\sum_k \sum_{i \in \boldsymbol{\tau}_p} \exp\{\mathbb{E}_q[\log \theta_{ik}] + \mathbb{E}_q[\log \beta_{kv}]\}} \tag{1.4}$$

I then calculate the contribution share of inventor i to patent p as follows

$$\omega_{ip} = \sum_{vk} \phi_{p,v,i,k}$$

This gives the probability that inventor i contributes a word to patent p, summing across all words $v \in V$.

1.2.1 Prior Versus Posterior Contributions

The Latent Dirichlet Allocation model outlined previously assumes a uniform prior over the contribution shares. In other words, each inventor in a team contributes equally ex-ante. Mathematically, this means that for any given word in patent p, the probability of selecting inventor i is given by $P(i_{pn} = i \mid \boldsymbol{\tau}_p) = \frac{1}{m}$. However, after observing the patent text and inferring the latent knowledge classes, the posterior probability $P(i_{pn} = i \mid \boldsymbol{w}_p, \boldsymbol{\tau}_p, \dots)$ becomes non-uniform. This shift is captured by the variational parameter $\phi_{p,v,i,k}$, which approximates the posterior distribution over inventor-knowledge class pairs. Summing over all words and knowledge classes we obtain an ex-post measure of contribution, captured by ω_{ip}.

The estimate for ω_{ip} updates the initial uniform prior based on the observed data. Holding the team fixed, if an inventor's patenting history, captured by their knowledge class distribution θ_i, aligns strongly with the patent's observed content, their inferred contribution ω_{ip} will increase. This can be seen through equation 1.4, in that both θ_{ik} and β_{kv} appear in the numerator of the definition for ω_{ip}. Therefore, while the prior assumes equal contribution shares, the posterior updates this given the knowledge content of the observed patent. This method allows the data to speak on who contributed more within the team by leaning on the powerful Bayesian logic behind the topic model implemented.

[7] I do so using the process outlined in the literature so again, leave the interested reader to consult Mortensen, 2017 for further details.

1.2.2 Measuring Concentrated Contribution Shares

I build a measure of the concentration of contribution shares to capture whether each team member contributes equally, or a few team members dominate. Specifically, I measure concentration as the Euclidean distance between the estimated vector of contribution shares across team members ω_p and a uniform distribution defined as $\bar{\omega}_m = \frac{1}{m}\mathbb{1}m$, where each member contributes equally. This metric captures the extent to which contributions are concentrated among a few individuals rather than being evenly distributed. A higher value indicates a more vertical team structure, where certain members dominate, while a lower value suggests a flatter distribution of contribution.

$$\text{Concentration}_{\tau p} = ||\bar{\omega}_m - \omega_p||^2 \tag{1.5}$$

1.3 Data

I use data on U.S. patents from the United States Patent and Trademark Office, collected through the repository *PatentsView*. This publicly available dataset provides the universe of U.S. patents, and their accompanying texts from 1976 onwards. They also provide a set of disambiguated inventor identifiers, allowing me to track team membership over patents. They also provide inventor characteristics including their gender and location.

1.3.1 Sample Selection

In order to infer a precise estimate of inventor and team level parameters, there are certain requirements that I make of the data. For example, long inventor patenting histories provide a richer set of data from which to learn the inventor's knowledge profile. However, most importantly, I require team switchers. The argument is similar to that made for the identification of an AKM fixed effect model (Abowd, Kramarz, and Margolis, 1999). By observing inventors appear on different teams, I can more accurately back out their knowledge profile and patent contribution. For example, consider the case of a computing specialist, who has written many computing patents with other computing experts. If they suddenly appear on a patent for a self-driving car alongside a transport expert, it is much easier to disentangle who contributed the automation knowledge compared to the engine structure.

I designed the following process to select a sample over which I can infer a precise contribution share. The sampling design prioritises teams that are made up of inventors who meet the two criteria of long histories, and regular team switching. Inventors and teams which patent only once are excluded, as their limited output provides little insight on inter- or intra-team knowledge structures. Similarly, inventors who have worked with only one team are removed to ensure that the final sample includes individuals who have contributed to multiple collaborations. Single-inventor teams are also dropped given that their contributions are immediate.

With this refined set of teams and inventors, I construct a bipartite network where nodes represent either inventors or teams, and edges capture the membership of inventors in teams. This network structure allows me to identify the most central teams, for which the ATM can precisely measure within-team contribution shares. I compute the betweenness centrality for all teams in the graph. The betweenness centrality allows me to prioritise teams composed of inventors who are common bridges across teams. By selecting teams with the highest centrality, the sample is composed of teams that contain inventors for whom I can get a precise estimate of both inventor and team level parameters.

Having selected the 10,000 most central teams, I extract the complete patenting history of the approximately 38,000 inventors who make up these teams. The resulting dataset consists of approximately 1 million patents.[8] This approach means that knowledge class distributions are inferred based on the entire career trajectory of inventors. Therefore, this is a static model, and the model does not allow for learning. An inventor's contribution to a given patent is estimated using a knowledge distribution learned from all their past and future patents.

1.4 Estimation

Pre-processing the patent texts is an important step to eliminate noise and ensure a good representation of the data. I first pre-process each patent p into a bag-of-words d_p. This bag of words contains only the informative words which characterise the knowledge contained in the innovation. The process begins by

[8]The fact that 38,000 inventors feature on approximately 1 million patents is in part a reflection that by selecting well connected teams, I will be selecting teams with prolific inventors. It is also a signal of the difficulty in balancing computational requirements and sample selection, since any small sample of inventors expands exponentially as you require the set of their co-inventors and their co-inventor's co-inventors etc.

converting all text to lower-case and filtering out punctuation. Each patent text is then tokenised to split it into individual words. Each token is stemmed, reducing words to their base form to consolidate variations of the same word, for example connect, connecting, connection, connected, all stem to *connect*. I remove a set of stopwords, which carry little semantic value due to their regular use in patents and scientific texts, alongside words shorter than three characters.[9]

TABLE 1.1

LDA PARAMETERS

K	η	Iterations	Passes	γ
50	1/K	350	100	0.001

Notes: K is the number of knowledge classes. η the Bayesian Dirichlet prior on the knowledge class to word distribution. Iterations sets the number of cycles used to update the knowledge class distributions, passes refers to the number of times the model goes over the entire dataset, and the gamma threshold sets the stopping point when the difference between topic updates is sufficiently small. The model has been run various times changing these parameters, and the results remain similar. Both η and γ are set to the `Gensim` default values. For more details consult the ATM package documentation online.

Table 1.1 provides the hyper-parameters which govern the estimation process. The number of passes defines the number of times that the model sees the entire dataset. The number of iterations defines the number of times the model iterates within the EM stage over each document. The model is trained using the online method where documents are loaded in batches of 2000. The choice of $\eta = 1/K = 0.02$ is the Gensim default option but also in line with the literature as both Hansen, McMahon, and Prat (2018) and Griffiths and Steyvers (2004) set $\eta = 0.025$. I estimate the Bayesian parameter flexibly, rather than setting a fixed prior. This allows for variation in the importance of a knowledge class on aggregate, which reflects a more natural state of the world.

Identification in a Bayesian context is not the same as in frequentist regression models, though there are similarities. If two inventors work together and produce many patents, but only ever working as a pair, it is impossible to disentangle who did what on those patents. In this case the model defaults to an equal probability for each team member across the knowledge classes contained within the patent. This is conceptually equivalent to assigning patent technol-

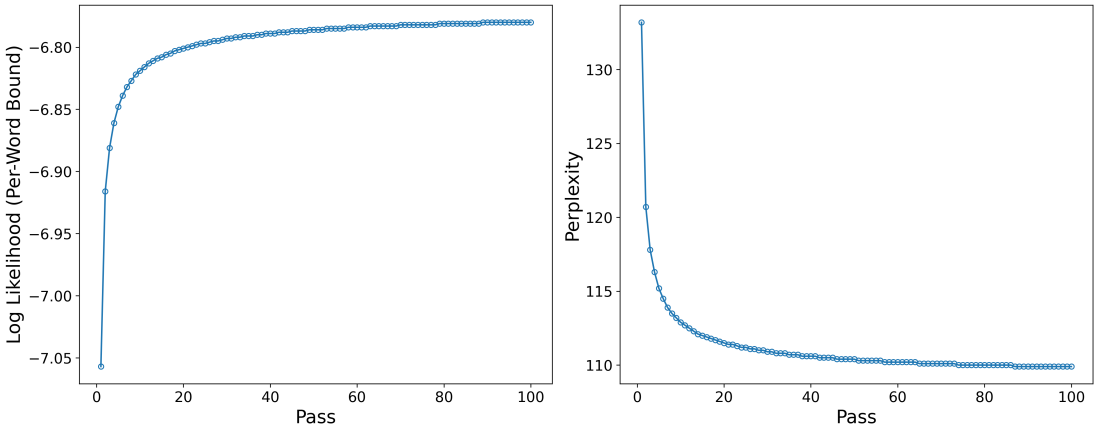
[9] I collect these stopwords from a range of sources, but they include common patenting and scientific words Sarica and Luo, 2020

ogy classes evenly across all team members (Jaffe, 1986). In addition, a topic model makes use of all documents fed into the model to identify the knowledge classes distributions, therefore even if the inventor level parameters are not well-identified, their patents still contribute to estimating other model parameters.

1.4.1 Convergence

The log-likelihood per-word bound and perplexity are common statistics used to measure convergence in topic models. Both statistics evaluate how well the model fits the data. The log-likelihood per-word bound measures the probability of the observed words given the estimated topic distributions, normalised by the total number of words. A higher bound indicates a better fit. When the value stabilises, the model has converged. Perplexity, which is the exponentiated negative log-likelihood per word, provides an interpretable measure of how uncertain the model is when predicting unseen data. Lower perplexity values correspond to a better model, as they indicate a more confident and accurate topic assignment. Figure 1.3 shows that the ATM converges in both these statistics.

FIGURE 1.3
LDA MODEL CONVERGENCE



Notes: Two convergence plots. One showing the log likelihood per word bound, and the second the model perplexity. Each statistic is calculated after every 25 iterations over the data.

1.4.2 Alternative LDA Parameters

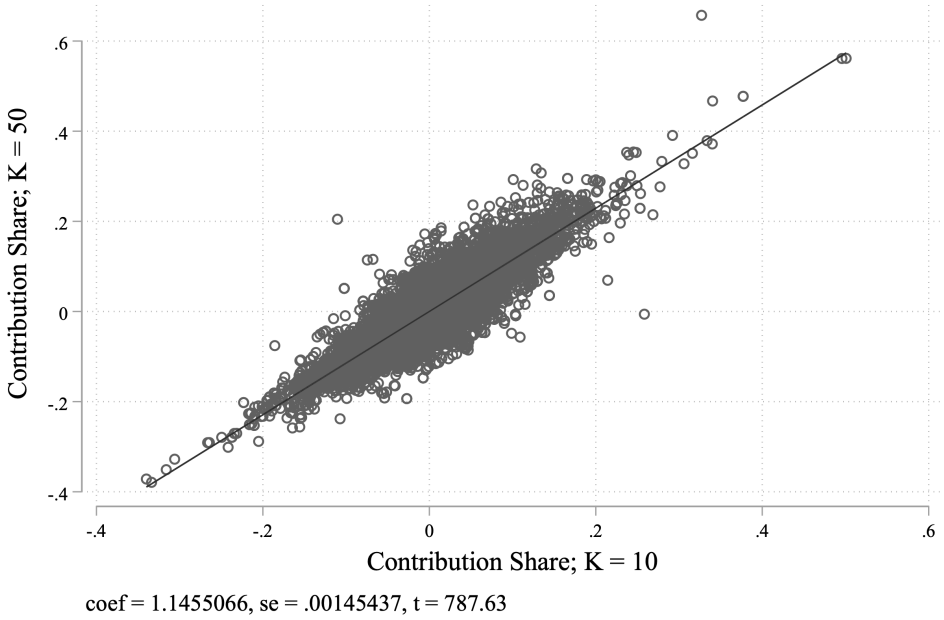
The key parameter chosen by the econometrician for this model is the number of knowledge classes K. Teodoridis, Lu, and Furman (2022) present a model in which they optimally back out the parameter K, and they find that $K = 77$ is optimal on a similar data sample of U.S. patents. Future research can look to

combine their approach with the one presented here. For now, I take that $K = 50$ to be a similar solution. Since an inventor's knowledge profile and contribution share are both continuous bounded variables, in theory, the choice of K is not a key determinant of the later empirical analysis.

Choosing K is most important to get a good representation of the text. I run the model for a K equal to 10 and 50. For the main analysis I proceed with the $K = 50$ model. In Figure 1.4, I show that the contribution weights for each inventor, conditional on a set of team size dummies, are highly correlated. The Pearson correlation coefficient, again conditional on team size, is 0.872.[10]

FIGURE 1.4

CONTRIBUTION SHARE FOR K = 10 & 50



Notes: A scatter plot between the contribution share inferred from a model with 10 topics compared to 50. All other LDA parameters remain constant. The plot contains a linear fit, and reports the regression model coefficient, standard error and t-score.

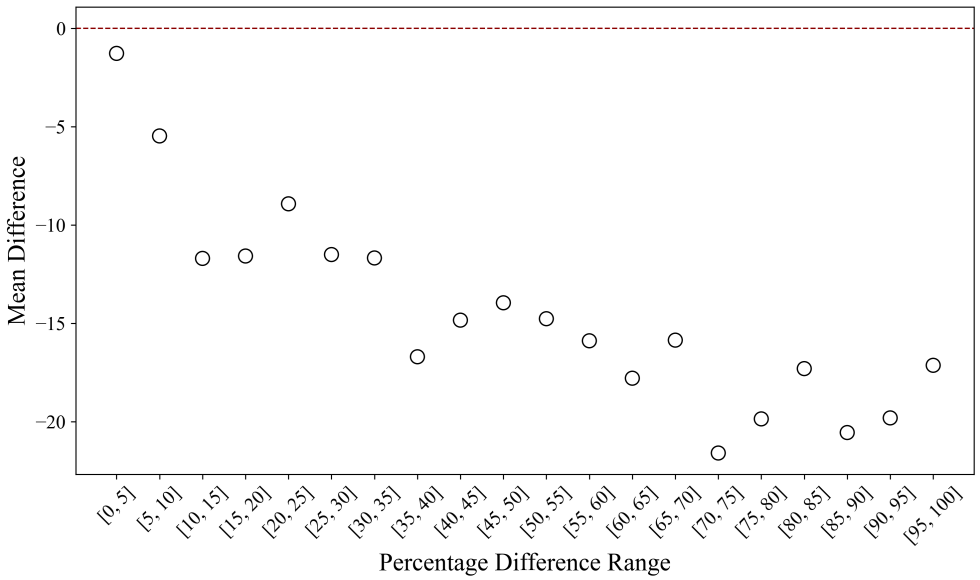
1.5 Validation

This paper is the first to estimate the contribution of each team member to the knowledge contained in a patent. To demonstrate the power of this method, I validate the inventor contribution weights using a prediction model. I propose that if the weights capture information on the true contribution share of each

[10] In the appendix, I present a further set of results comparing across values of K.

inventor to a patent, then the patenting history of inventors who contribute significantly more should be a stronger determinant of the technology classification awarded to the same patent.

FIGURE 1.5
CONTRIBUTION SHARE VALIDATION



Notes: A plot of the mean difference between the feature importance for either the lead or second inventor on a patent, when predicting a patent's CPC classification. The lead and second inventor are determined from the inferred contribution weights. The prediction model is random forest. For each bin I run 100 different splits of the data. This is a form of cross-validation that removes the dependency of the outcome on a random initial seed and allows me to estimate a standard error, however they are very small and uninformative in the figure.

For each patent in the sample I define the lead and second inventor by ordering their estimated contribution shares and calculate the percentage difference between them. I use a random forest to predict the CPC classification awarded to a patent with two sets of explanatory variables: the five most common CPC classes used by the lead inventor, prior to the target patent, and the corresponding five for the second inventor. When using a random forest you can then calculate the feature importance for each explanatory variable, similar in concept to measuring how each variable contributes to the R^2 of a regression.

I propose that if the gap between the contribution shares of the two inventors is large, then the lead inventor's patenting history will be a significantly stronger predictor of the CPC class awarded to a patent. While if that difference is small, then I predict there to be no significant difference. This corresponds to the total feature importance for the lead inventor's patenting history being significantly larger than that of the second inventor.

Figure 1.5 plots the difference in feature importance between the first and second inventor's patenting histories, across binned groups of the difference between their inferred ATM-contribution share. The feature difference is defined as the feature importance for the second inventor minus that of the lead inventor. Therefore the hypothesis is that this difference is negative, and decreasing further as the percentage difference between the contribution share of the lead to second inventor increases. There is a strong negative trend, showing that the greater the difference in the inferred ATM contribution share, the more information the leading inventor's patenting history provides in the prediction task. This points to the contribution weights providing economically important and precise information on who contributed to the knowledge contained.

1.6 Descriptive Results

Table A.1 presents a set of descriptive statistics on the core sample of team- and individual-level parameters. Given that the sample is selected on the teams recording the highest centrality within a bipartite inventor to team network, there may be valid selection concerns. For example, the average team size is far larger in this sample than for the universe of patents. By selecting the most central teams, the method oversamples from larger teams as they are more likely to connect team nodes through prolific inventors. Therefore, the results presented here should be considered a proof of concept demonstrating how this framework can provide novel statistics on collaboration patterns.

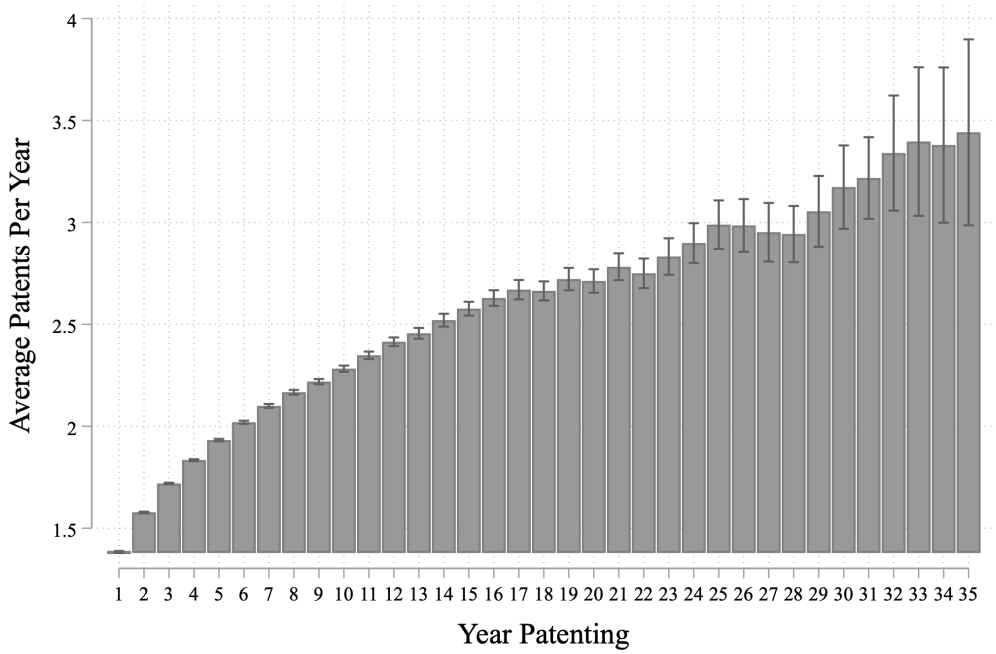
1.6.1 Experience and Contribution Shares

Consider the following empirical fact shown in Figure 1.6. The number of patents an inventor collaborates on in that year increases with each additional year of patenting experience. This result comes from an inventor fixed-effect regression, where Y_{it} is the number of patents inventor i appears on in year t. The sample is the universe of USPTO patents from 1976 to 2024. The regression includes a set of dummy variables capturing whether inventor i in year t has exactly s years of patenting experience. I also include year fixed effects (δ_t). Formally, the regression is:

$$Y_{it} = \alpha_i + \sum_{s=1}^{35} \beta_s \cdot \mathbb{1}(\text{years experience}_{it} = s) + \delta_t + \epsilon_{it} \tag{1.6}$$

FIGURE 1.6

PATENTS PER YEAR OF EXPERIENCE



Notes: Plots the average number of patents that an inventor collaborates on, in each year of their patenting career. This is taken from the predicted values of a regression of the number of patents an inventor collaborates on in one year, on an inventor fixed effect and year dummies. The sample taken is the universe of USPTO patents from 1976-2024.

Does this reflect an increase in productivity with experience? To answer this question I examine how much each inventor i contributes to patent p through their contribution share ω_{ip}. In order to examine relative contribution within a team of m inventors, and to facilitate comparison across team sizes, I define the outcome variable as a normalised contribution share

$$Y_{ip} = \frac{\omega_{ip}}{1/n_{\tau}}$$

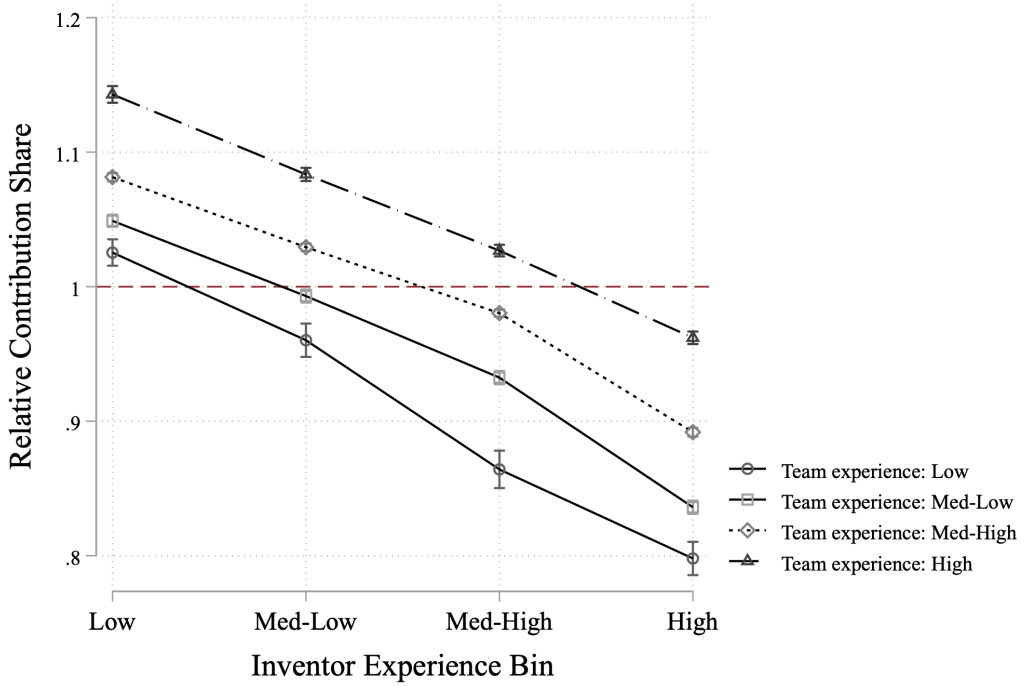
I measure patenting experience as the cumulative count of the number of patents that inventor has produced, prior to patent p. I split this variable into 4 equally sized bins to track the experience of inventor i on patent p ($exp_{i(p)}$). These refer to low, medium-low, medium-high and high experience levels. I build a second count for the mean number of patents the inventor's collaborators in team τ have produced, prior to the patent p. I denote the team τ minus inventor i by $\tilde{\tau}$. I split this same variable into the same four bins ($exp_{\tilde{\tau}(p)}$). This allows me to track whether inventor i collaborated with junior or senior co-inventors.

I estimate how an inventor's contribution share varies with their experience level relative to their co-inventors through a team level fixed-effect regression. I control for unobserved heterogeneity at the team level through a team fixed effect α_τ. I include indicator functions for inventor experience, average team-mate experience, and their interaction. I control for CPC technology class indicators (δ_c), and year fixed effects (δ_t). I cluster standard errors at the team level.

$$Y_{ip} = \alpha_\tau + \sum_{e=1}^{4} \beta_e \cdot \mathbb{1}(\text{exp}_{i(p)} = e) + \sum_{o=1}^{4} \beta_o \cdot \mathbb{1}(\text{exp}_{\tilde{\tau}(p)} = o) + \sum_{e=1}^{4}\sum_{o=1}^{4} \delta_{eo} \cdot \mathbb{1}(\text{exp}_{i(p)} = e) \times \mathbb{1}(\text{exp}_{\tilde{\tau}(p)} = o) + \delta_c + \delta_t + \epsilon_{ip}$$

FIGURE 1.7

CONTRIBUTION SHARE OVER EXPERIENCE



Notes: This figure plots the relative contribution share for an inventor, across each of the four quartiles of inventor experience. Team experience is defined as the mean experience of each of the inventor's co-inventors, removing the focal inventor. This count is also split into four bins, using the same thresholds as defined by the inventor experience quartiles. The Y-axis gives the predicted relative contribution share from a regression including team, year and technology class fixed effects. Standard errors are clustered at the team level.

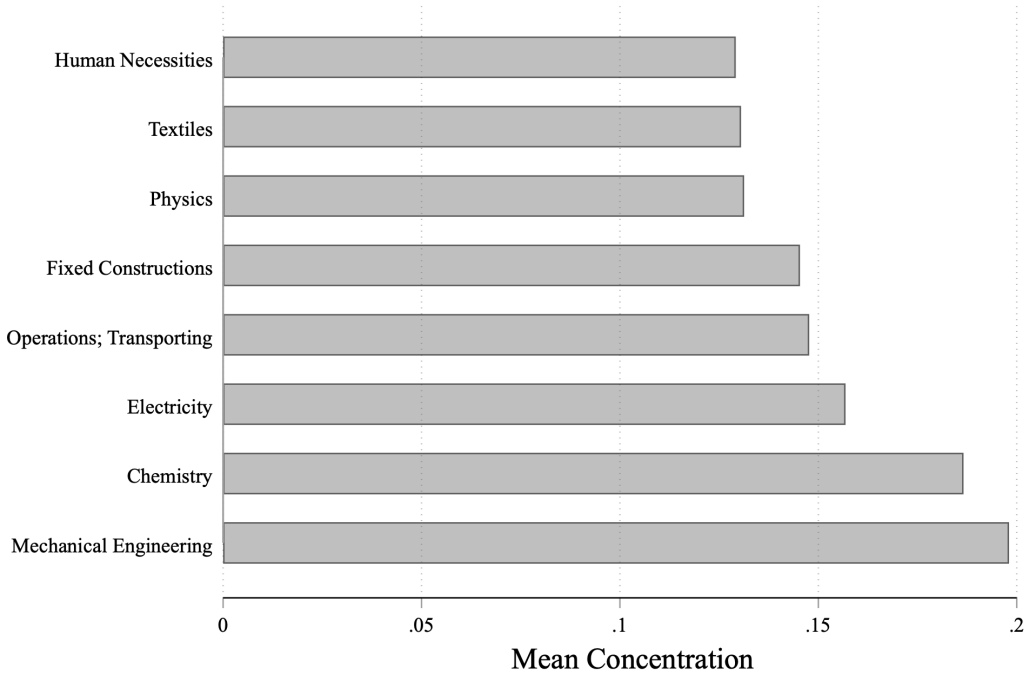
Normalising by a uniform contribution share, Figure 1.7 allows me to show

multiple trends simultaneously which together reveal interesting within-team collaboration dynamics. I define senior inventors as those with high levels (top quartile) of experience, and juniors those with low levels (bottom quartile).

When collaborating with junior co-inventors, senior inventors contribute relatively less than their junior colleagues. Combined with the result from Figure 1.6, this points to a diluting effect, where senior inventors collaborate on more projects, but do less on each one. Interestingly however, this effect disappears when senior inventors collaborate with other seniors, as their relative contribution share tends to 1. For junior colleagues, again, the same trend holds. When a junior collaborates with other juniors their relative share is approximately 1. This indicates that on average they each contribute equally. This points to the importance of *relative* seniority in determining the scientific division of labour within teams.

FIGURE 1.8

CONCENTRATION OVER TECHNOLOGY AREA



Notes: This figure presents the mean concentration of inventor teams across CPC technology classes, where concentration is measured by the Euclidean distance between the vector of contribution shares and a uniform distribution. Bars are sorted by mean concentration, with labels corresponding to CPC sections. CPC titles are abridged.

I extend these results in appendix section A.3 to introduce inventor gender as a source of heterogeneity. I show that when you condition only on the inventors experience level, female inventors contribute more than their male co-inventors. This result holds for gender-diverse teams : a team that includes at least one female and one male. Further examination however reveals that this result is not driven by gender, but also by relative experience level. When conditional on the interaction of an inventor's experience level, and that of their co-inventors, this gap disappears. This points to female inventors having a larger contribution share, when collaborating with men, as they tend to be junior females collaborating with senior males. The same relative experience channel drives any observed differences across gender.

1.6.2 Concentration over Technologies and Time

I present a set of aggregate descriptive statistics for the concentration measure introduced in equation 1.5. Figure 1.8 demonstrates that teams divide contribution shares differently across technology classes. I plot the average concentration by CPC classification class. I find that contributions in Mechanical Engineering are, on average, more likely to be concentrated on a few team members. While Physics, Textiles, and Human Necessities (Medicine) tend to be more evenly distributed. These differences may be driven by a number of factors such as different levels of capital intensity or labour supply and warrant further research.

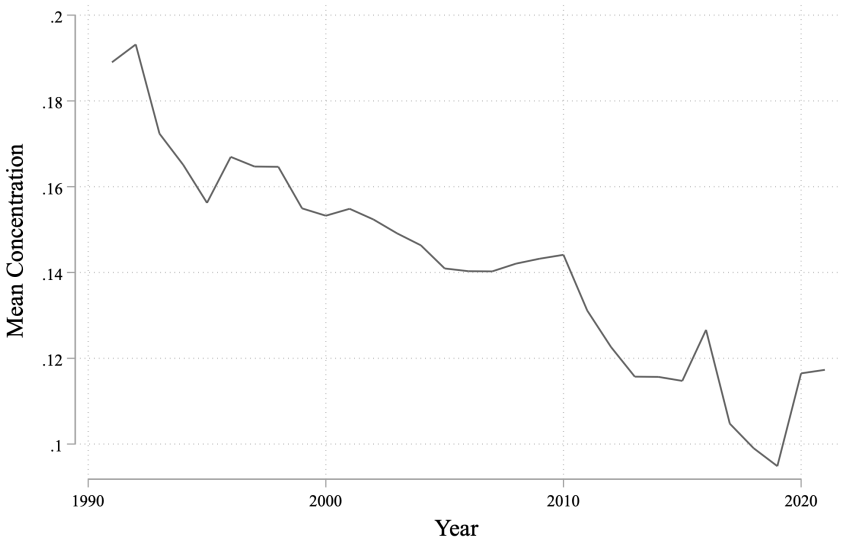
I then consider the time dimension. In Figure 1.9 I plot the average concentration across the patents produced in that year. This result shows that on aggregate, teams are getting flatter.

1.7 Main Results

Concentration varies over technology class, time, and is in part determined by the relative seniority of team members. The question is then what impact the division of contribution shares within teams has on patent outcomes? Xu, Wu, and Evans, 2013 first argued that scientific teams in which team members contributed equally produced more disruptive science and technology. I revisit that result in a new innovation context.

I measure patent value Y_p along five dimensions. I measure the number of cita-

FIGURE 1.9
CONCENTRATION OVER TIME



Notes: This figure presents the mean concentration of inventor teams over time, measured by the Euclidean distance between the vector of contribution shares and a uniform distribution. The sample is taken from 1991-2021.

tions received using USPTO data.[11] I measure the market value, in millions of USD, for the sub-sample of patents awarded to public US firms, using data from Kogan et al. (2017). I measure whether a patent is a breakthrough using data provided by Kelly et al. (2021). The breakthrough measure is a binary outcome, derived from comparing the similarity of the text to patents that came before and those that came after. This concept of breakthrough captures whether that patent started a new research area. Finally, I measure a patent's novelty and impact using data from Arts, Hou, and Gomez (2021). Novelty is measured through the number of new words that a patent introduces to the USPTO lexicon. Impact is measured by the number of new words, weighted by how many times they are reused by future patents.[12] These dimensions are summarised in the following table.

[11] I don't normalise by technology class year trends at this point, as I will introduce both technology class and year fixed effects to the regressions.

[12] Taken directly from the Read Me file for Arts, Hou, and Gomez (2021): "For patent p, new_word_reuse$_p$ $= \sum_{i=1}^{n}(1 + u_i)$, where n is the number of new keywords introduced by patent p and u_i is the number of future patents that reuse the new keyword i"

TABLE 1.2

SUMMARY STATISTICS ON PATENT OUTCOMES

#	Outcome Variable	N	Mean	Std. Dev.	Skewness
1	Citations	27,127	16.547	46.193	14.768
2	Market Value	14,922	3.810	20.146	26.817
3	Novelty	26,762	3.724	17.933	26.857
4	Impact	26,762	16.212	92.981	32.468
5	Breakthrough	20,129	0.130	0.337	2.194

Notes: This table presents a set of summary statistics on the five patent outcomes. These outcomes are sourced from external data sources: (1) PatentsView, 2024; (2) Kogan et al., 2017; (3) and (4) Arts, Hou, and Gomez, 2021; (5) Kelly et al., 2021.

I first run an OLS regression, given by equation 1.7, including the same technology class, team size and year dummies as previous.[13][14]

$$Y_{\tau(p)} = \beta_0 + \beta_1 \text{concentration}_{\tau(p)} + \delta_c + \delta_m + \delta_t + \epsilon_{\tau(p)} \tag{1.7}$$

The coefficients are large, partly due to the units of the concentration measure. Therefore, these results are best understood in terms of standard deviations. A one standard deviation increase in concentration leads to a 17.6% decrease in citations received by a patent; a 34.4% decrease in the patent's market value; a 5.7% decrease in its novelty and a 5.6% decrease in its impact. Finally, the probability of producing a breakthrough innovation declines by 5.8 percentage points when concentration increases by one standard deviation. Importantly for interpretation, each coefficient is negative and highly significant, which backs up the claim that flat teams drive science.

The results on the market value and citations received remain large, and in part

[13] I do not include a team fixed effect here, since variation in concentration is largely between teams, not within. The standard deviation of concentration across all patents is 0.0914, whereas the mean within team standard deviation is 0.00887. The ratio of mean within team SD to the overall SD is approximately 0.097. In other words, less than 10% of the total variation in concentration comes from within team variation.

[14] One valid potential bias is that teams made up of inventors who patent only a few times will tend to have more equal contribution shares. This is because the LDA cannot distinguish between each inventor's contribution due to short patenting histories. Having few patents may reflect lower quality inventors, which will therefore correlate with poorer patent outcomes. Therefore, as a robustness check I introduce a control for the total and mean number of patents awarded to the team inventors. The results do not change and are available upon request.

can be explained by the magnifying effect of the distribution of the concentration measure and each outcome variable. Table 1.2 shows that citations, and particularly the market value have very strong variance and are highly skewed, which combined with a highly variable and skewed concentration measure amplifies the estimated effect.

While novelty and impact also both show significant variance and skewness, the estimated coefficients suggest that the underlying relationship is weaker; therefore the amplification effect is smaller. The point estimates may require further examination, but the trend is clear, concentration tends to correlate with lower value patents.

TABLE 1.3

CONCENTRATION ON PATENT OUTCOMES

	(1) ln(Citations+1)	(2) ln(Market)	(3) ln(Novelty+1)	(4) ln(Impact+1)	(5) Pr(Break)
Concentration	-1.9334***	-4.6084***	-0.6339***	-1.0018***	-0.2366***
	(0.0844)	(0.2161)	(0.0629)	(0.0950)	(0.0229)
N	27103	14917	26738	26738	20111
R^2	0.323	0.227	0.189	0.219	0.194

Notes: This table presents regression estimates examining the relationship between team concentration and five innovation outcomes: citations, market value, novelty, impact and the likelihood of producing a breakthrough patent. Concentration is measured as the Euclidean distance between the vector of contribution shares and a uniform distribution. All models include year fixed effects, and robust standard errors are used.

1.7.1 Junior and Senior Quality Effects

These results should be interpreted in the context of the empirical facts demonstrated previously. Concentration is driven largely by senior inventors contributing less, and their junior colleagues doing relatively more. This can be interpreted as the seniors increasing their quantity-to-quality trade-off. Seniors collaborate on more patents. However, when collaborating with more junior colleagues, they contribute less to each patent. If these junior colleagues are lower quality, this will drive patent outcomes.

To examine this effect further, I define a simple measure for inventor quality as a weighted average of the patent outcomes that inventor earns throughout their career.[15] For these weights I use their contribution share to the corresponding

[15]If the current patent p is dropped from this sum, for concerns that it is driving the outcome

patent. For an inventor working on the set of P_i patents, I define inventor quality for each patent outcome Y as

$$\text{quality}_i^Y = \sum_{p \in P_i} \omega_{ip} Y_{\tau(p)}$$

This quality measure assumes a constant inventor quality, that is revealed throughout the inventor's career.[16] I then define this at the team $\times$ patent level as the quality of the lead inventor. Where the lead inventor is defined as the inventor with the largest contribution share on patent p.

$$\begin{aligned} Y_{\tau(p)} = \beta_0 + \beta_1 \text{concentration}_{\tau(p)} + \beta_2 \text{lead quality}^Y_{\tau(p)} \\ + \beta_3 \cdot \text{concentration}_{\tau(p)} \times \text{lead quality}^Y_{\tau(p)} + \delta_c + \delta_m + \delta_t + \epsilon_{\tau(p)} \end{aligned}$$

Both concentration and inventor quality are continuous variables. Table 1.4 shows that if the lead is higher quality, then naturally there is a boost to the corresponding patent outcome. This is shown through the positive and significant coefficient on lead quality, and simply reflects that they have increased the average quality within the team. Most important however is that the interaction term between the quality of the lead inventor and the concentration measure is significant and positive, for all patent outcomes. So while concentration remains negatively correlated with all patent outcomes, this effect is mitigated by the quality of the lead inventor.

In Figure 1.10, I zero in on the impact of collaborating with high quality juniors. I split the regression coefficient further with a three-way interaction. I interact the inventor-team experience interaction with an indicator for the lead being a junior. I condition on the lead inventor being in the lowest experience bin, which corresponds to having fewer than 6 previous patents.

In Figure 1.10, I predict the number of citations that a patent will receive over different levels of concentration, split over the quality of their junior inventor.[17] We clearly see that for low quality inventors, concentration reduces the number of citations received. This is a natural result, since the lower quality inventor is over-represented in the knowledge contained in the patent. Converting the Y-

measure directly, the results hold.

[16] If instead the sum was taken over patents produced prior to patent p, this measure would simply capture the fact that they are junior. For seniors to identify quality juniors, they would need to predict future outcomes, or identify a constant but not yet revealed quality.

[17] Given the coefficient estimates in Table 1.4 the same graph for each of the other four outcomes would be very similar.

axis from its logarithm scale shows that moving from low to high concentration decreases citations by almost 40%, when the lead inventor is both junior and low quality. However, when the lead is a high quality junior, this effect is mitigated entirely.

TABLE 1.4

CONCENTRATION AND LEAD QUALITY ON PATENT OUTCOMES

	(1) ln(Citations+1)	(2) ln(Market)	(3) ln(Novelty+1)	(4) ln(Impact+1)	(5) Pr(Break)
Concentration	-1.7693***	-3.4897***	-0.7392***	-1.0770***	-0.2959***
	(0.1023)	(0.1937)	(0.0682)	(0.1026)	(0.0207)
Lead Quality	0.0003**	0.0486***	0.0047***	0.0015***	0.0221**
	(0.0001)	(0.0032)	(0.0009)	(0.0003)	(0.0084)
Concentration × Lead Quality	0.0105***	0.1106***	0.0520***	0.0147***	0.7829***
	(0.0013)	(0.0318)	(0.0073)	(0.0023)	(0.0686)
N	27103	14917	26738	26738	20111
R^2	0.375	0.396	0.209	0.237	0.252

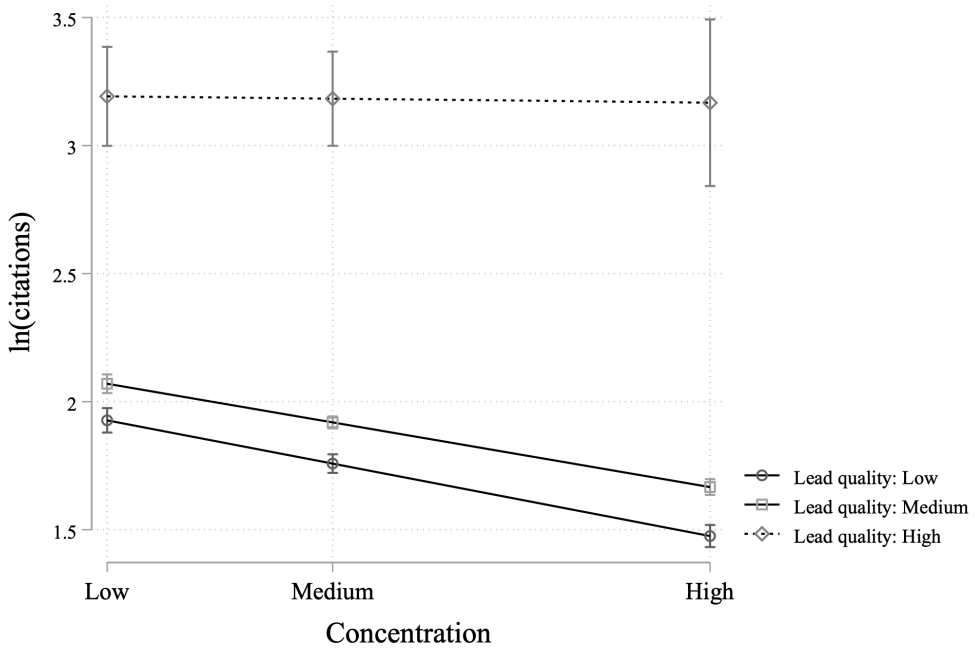
Notes: This table presents regression estimates examining the relationship between team concentration and five innovation outcomes: citations, market value, novelty, impact and the likelihood of producing a breakthrough patent. Concentration is measured as the Euclidean distance between the vector of contribution shares and a uniform distribution. All models include year fixed effects, and robust standard errors are used.

It is natural to suggest that the quality of the senior inventor may also determine patent outcomes. A high quality senior, who has taken on a potentially more backseat role, may guide the junior members to better outcomes. Table 1.5 presents the results from a synonymous regression, however now interacting the quality of the senior with the concentration measure. A senior is identified as the inventor with the highest experience level when producing patent p.

$$\begin{aligned} Y_{\tau(p)} = \beta_0 + \beta_1 \text{concentration}_{\tau(p)} + \beta_2 \text{senior quality}^Y_{\tau(p)} \\ + \beta_3 \cdot \text{concentration}_{\tau(p)} \times \text{senior quality}^Y_{\tau(p)} + \delta_c + \delta_m + \delta_t + \epsilon_{\tau(p)} \end{aligned}$$

Interestingly, while again the quality of the senior leads to a jump in levels, it does not appear to mitigate the role of concentration. Except for the market value of a patent, and a small effect on their ability to create a breakthrough. This emphasises that when a senior inventor changes their role within a team, if collaborating with junior members, patent outcomes are driven more by the quality of the juniors they collaborate with, than their own innate quality.

FIGURE 1.10
CONCENTRATION × LEAD JUNIOR QUALITY ON CITATIONS



Notes: A plot of the predicted natural logarithm of citations a patent receives, over the concentration within the team, split over three levels of quality, low, medium and high. This plot shows the effect for junior leads. Both inventor quality and concentration are measured as continuous variables, and low, medium and high are defined as the 10^{th}, 50^{th} and 90^{th} percentile.

TABLE 1.5
CONCENTRATION × SENIOR QUALITY ON CITATIONS

	(1) ln(Citations+1)	(2) ln(Market)	(3) ln(Novelty+1)	(4) ln(Impact+1)	(5) Pr(Break)
Concentration	-1.8025***	-3.3894***	-0.5492***	-0.9779***	-0.2415***
	(0.0956)	(0.1973)	(0.0826)	(0.1218)	(0.0252)
Senior Quality	0.0007***	0.0332***	0.0030***	0.0008***	0.0249***
	(0.0001)	(0.0015)	(0.0005)	(0.0002)	(0.0038)
Concentration × Senior Quality	0.0004	0.0715***	-0.0031	0.0002	0.0529*
	(0.0005)	(0.0136)	(0.0029)	(0.0009)	(0.0221)
N	27103	14917	26738	26738	20111
R^2	0.334	0.394	0.192	0.223	0.203

Notes: This table presents regression estimates examining the relationship between team concentration and five innovation outcomes: citations, market value, novelty, impact and the likelihood of producing a breakthrough patent. Concentration is measured as the Euclidean distance between the vector of contribution shares and a uniform distribution. All models include year fixed effects, and robust standard errors are used.

These results demonstrate that if seniors delegate contributions to junior co-inventors in order to split their time over more individual patents, they face a potential quantity-quality trade-off. A senior can increase the total quantity of patents by delegating, but if they concentrate the share onto low quality co-inventors then the quality of the patent as a whole reduces. This is backed up in the data through the negative correlation between concentration and value. One potential outcome could be that the senior makes smaller but more precise, and therefore equally as important contributions. I present supporting evidence that the quality of the senior does not make up for the loss of concentrating contributions onto lower quality team members.

1.8 Conclusion

This paper introduces a novel method for modelling team collaboration that allows for the scientific division of labour. The statistical model, when leveraged to high-dimensional patent text data, can disentangle individual contributions to the knowledge production process. I apply the model to demonstrate a new set of descriptive statistics on within-team organisation of labour.

By backing out a contribution share, this paper contributes to our understanding of whether flat teams produce better science. I demonstrate that as inventors become increasingly experienced they collaborate on more patents each year; however contribute less to each individual patent. In addition, when seniors collaborate with juniors, the juniors have a larger contribution share. This fact can explain the negative correlation between concentration and patent outcomes, since I show that if the junior is low quality then indeed concentration is correlated with lower patent value. However, this effect is mitigated entirely if the junior is of high quality. This opens the door to a wealth of future research on the direction of technological change, allowing us to better understand the role of personal and inter-personal dynamics in driving innovation.

Chapter 2

Catalyst or Constraint: The Dual Role of Prior Innovation for Breakthroughs

Abstract

This chapter studies the impact of an expanding scientific and technological frontier on team innovations. I model collaboration directly through a Bayesian model of Natural Language Processing. Applied to patent text data, this model builds a map of inventors, teams, and research fields, referred to as the *knowledge space*. Applied to over 2.2 million U.S. patents from the USPTO, this framework allows me to tackle unanswered questions on how teams create new knowledge. Specifically, I investigate the effect of prior work on a team's ability to produce a breakthrough—an innovation that sparks a new and successful research field. Leveraging high-dimensional patent text data, I back out two new measures: breakthrough patents and a team's knowledge field. I combine this with data on premature inventor deaths as a quasi-natural experiment. This identifies how team innovations change as they pivot to more or less advanced research fields. Teams build on existing knowledge, and prior work both supports and obstructs innovation. I show that teams generate more breakthroughs when building on enough prior work to incorporate valuable knowledge, but not so much as to stifle novelty.

I gratefully acknowledge the support of the Spanish Agencia Estatal de Investigación (MCIN/ AEI /10.13039/501100011033) through grant PID2020-114251GB-I00. This project was partly developed while visiting the University of California, Merced, for whom I also give thanks for their support and guidance.

2.1 Introduction

Organising inventors into effective teams is essential for technological growth but also for addressing society's greatest challenges. Literature suggests that the dominance of teamwork is partly driven by an ever-increasing knowledge stock (Jones, 2009), as growing fields present increasingly complex problems. However, this prior literature has largely focused on innovation value through citations, overlooking how teams contribute to the creation of new and successful research fields.[1]

In this paper, I study the impact of an expanding scientific and technological frontier on team innovations. To do so, I present a novel framework that integrates inventor teams and their patent texts. I model collaboration directly through the lens of a Bayesian model of Natural Language Processing (NLP). Applied to patent text data, this model builds a map of inventors, teams, and research fields, referred to as the *knowledge space*. Leveraging high-dimensional patent text data and a tractable model of collaboration, this framework allows me to answer questions on which systematic data was missing from the literature. Specifically, I study the impact of prior work on a team's ability to produce a breakthrough—an innovation that sparks a new and successful research field. Given this, I find that teams produce more breakthroughs when building on enough prior work to incorporate valuable prior knowledge; however, not too much that it becomes hard to be novel.

The analysis in this paper proceeds in two steps. I first develop a method to characterise the latent knowledge held by inventors and their patents, disentangling the individual contribution of each team member. I train the model on 408,774 U.S. patents from 214,535 teams using the USPTO *PatentsView* database. Using the trained model, I fit an additional 2.2 million U.S. patent texts into the knowledge space. This novel space allows me to back out two new empirical measures. First, a breakthrough patent is an innovation in a research field with little prior work, which afterwards grew into a vibrant research area. Second, a team's knowledge field, defined as the set of all research fields accessible to the team. In the second stage, I combine this with data on premature team member deaths (Kaltenberg, Jaffe, and Lachman, 2021). This provides a quasi-random

[1] Key references studying teamwork and knowledge production through citations include Pearce, 2022; Bonhomme, 2022; Ahmadpoor and Jones, 2019, consult the literature review for a more detailed discussion.

shock to a team's knowledge field. Through a continuous treatment model, I identify how team innovations change as they pivot to more or less advanced research fields.

I document that research fields have become increasingly crowded over time, which has meaningful consequences for whether teams achieve breakthroughs. I find that the likelihood of a patent sparking a breakthrough follows an inverted-U shape with respect to prior work. Building on some prior work boosts innovation impact, but too much stifles novelty. This translates directly to team outcomes. For teams in advanced fields, removing a member from more established areas and shifting focus to less-explored fields increases their breakthrough potential by as much as 50%. I show that this is driven by these teams being more novel, as they introduce more new words to the USPTO lexicon. The opposite occurs for teams in early-stage fields. Reducing the quantity of prior work on which these early-stage teams can build reduces the chances that their next patent sparks a breakthrough. I present evidence that these teams, however, do not become more novel, and that this reduction is driven by a fall in impact. They are in fact on the upward sloping region of the inverted-U shape, and would do better by incorporating more established knowledge.

I contribute to the literature by constructing a unifying framework for studying teams. This framework not only consolidates existing insights but also broadens the scope of team research beyond traditional metrics.[2] Previous studies have largely focused on measuring team value through citations and examined team composition using low-dimensional categories of inventor types. These methods are less appropriate to study the creation of new research fields. This paper addresses these limitations. By combining high-dimensional patent data with a statistical model of teamwork, I develop a new method to disentangle each team member's contribution to a patent's knowledge content. By leveraging the high dimensional patent text data the model can back out a representation of the team in data. This allows me to locate the team in the knowledge space, describe the research areas on which they could build, and characterise how much prior work existed in those areas. This represents a contribution to our ability to study the production of breakthrough innovations, which had often be neglected due to a lack of data or models capable of capturing the creation of

[2]To demonstrate its effectiveness, I replicate two well-known findings on team composition and breakthroughs: small teams are the most disruptive, and flat teams drive radical science (Xu, Wu, and Evans, 2013; Wu, Wang, and Evans, 2019).

new scientific fields.

I develop and apply a two-part empirical strategy to demonstrate the results. First, I represent the latent knowledge held by inventors and patent texts. I model a patent as a combination of knowledge classes. Each class represents a specific domain of expertise. For example, a car includes knowledge on engines, wheels, fuel, etc., and some on computing. The first self-driving car then increased the amount of computing knowledge in order to automate driving. The first patent to do so was novel, but what can explain why this became a breakthrough research area? I define the knowledge space as a probability simplex across a set of knowledge classes. Inventors are characterised by their position in this space. Teams innovate by combining the knowledge profile of each member. A team's knowledge field is then defined by all possible combinations of its members. This corresponds to the set of research fields available to the team. Through a simplex, this approach naturally incorporates a spatial concept by embedding a notion of distance. I can then measure the development stage of each research field available to the team by counting the amount of prior work in each area of the knowledge space. I show in this paper that the quantity of prior work a team builds on indeed explains which novel ideas become breakthroughs.

A premature death, defined as the death of an active collaborator, serves as a random shock to a team's local knowledge field. The use of premature deaths as a source of exogeneity is well established (Azoulay, Fons-Rosen, and Graff Zivin, 2019; Azoulay, Graff Zivin, and Wang, 2010). The death of a collaborator changes the set of research fields available to the team by removing potential combinations. I apply a continuous treatment model to show that a team's innovation output is determined by the prior work in this new set of research fields. I start by showing that the premature death of a team member leads to changes in the knowledge content of a team's innovation, as revealed by the language in the patent text.

The impact of a death on a team's research depends on which team member is lost and their contribution to the team's local knowledge field. The average treatment is the mean decrease in the quantity of prior work in a team's knowledge field after a premature death. I predict how this change determines a team's ability to achieve a breakthrough. Following the death of a team member, the average treatment increases the likelihood of producing a breakthrough

by 21.27% relative to the baseline.[3] However, this result hides significant heterogeneity. When I split the sample over four quartiles of the quantity of prior work in a team's knowledge field, prior to the premature death, I find important heterogeneity in the treatment effect. For teams building on advanced areas, reducing the quantity of prior work by the average treatment increases their chances of a breakthrough by 49.7%. However, for those already working in early-stage research fields, the same change reduces their chances of a breakthrough by 61.4% on average.

The results presented here can be understood through the lens of an endogenous growth paradigm. Prior work exerts opposing effects on breakthroughs. On the one hand, prior work lowers the cost of innovating by providing a solid foundation. This aligns with the idea that moving up the quality ladder of development reduces implementation costs (Grossman and Helpman, 1991). However, when the goal is to create a new research field, prior work becomes an obstacle. It not only prevents teams from being the first to develop an idea but also establishes paradigms that shape future work. This relates to the literature on the burden of knowledge, and how the expanding scientific frontier is driving the rise in teamwork (Jones, 2009; Agrawal, Goldfarb, and Teodoridis, 2016). As the frontier of knowledge expands, inventors must invest more effort to develop on that frontier. At an aggregate level, as the knowledge space fills up, breakthrough ideas become increasingly difficult to find (Bloom et al., 2020).

These findings provide guidance for policymakers. Research funding should be distributed across fields, as concentrating it in one area may obstruct breakthroughs. Diversifying funding across new and advanced fields will help teams combine ideas in novel ways and foster breakthroughs. In addition they promote the use of cross-field collaboration. For teams working on advanced fields, by searching for new team members in up-and-coming fields they can find the novelty they need to spark a new and successful field.

On a technical level, this paper makes a contribution to the use of NLP models in economics. Patents have been a valuable proxy of innovation for decades, and this paper forms part of a growing literature making use of the depth of knowledge contained in their texts. Through a hierarchical Bayesian model, I infer who contributed which section of a patent text. Over each inventor's

[3]This number is calculated using the predicted values from the regression model. The average number of patents lost from the death of an inventor is 174.22, the baseline probability of a breakthrough is 0.44, given the coefficient 0.0022.

entire patenting history, the model learns their individual knowledge profile. If an inventor has a long history of producing AI patents and appears on a patent for a self-driving car with an inventor with a background in engineering, the model can distinguish between their contributions. It identifies who provided the knowledge on automation and who contributed to the engine structure. This highlights the key contribution of this method beyond patent technology classes. Patent classification systems provide an accurate description of the knowledge contained within a patent, however they do not provide enough data or variation by which to back out an individual inventor's contribution.

As a novel method, I validate this space along various dimensions by comparing the model to existing data. The model develops a measure of breakthrough patents as those that experience the largest growth in the number of patents within their research field, following their publication. Patents that I identify as breakthroughs introduce 8.67% more new words and 47.6% more new combinations of two existing words, which are subsequently reused by future patents.[4] This evidence reflects the paper's central premise: breakthrough innovations arise from the recombination of existing ideas.

Related Literature The first broad literature that this paper contributes to is on the importance of teams within science and technology. It is now taken as standard that teams are the principal producers of innovation (Wuchty, Jones, and Uzzi, 2007). A range of reduced form papers have looked to describe team composition and its role in explaining innovation outcomes (Uzzi et al., 2013; Xu, Wu, and Evans, 2013; Wu, Wang, and Evans, 2019). I present here a unifying framework for teamwork that replicates a selection of these results in one model.

There is a growing literature using individual wage data to explain productivity differentials and complementaries between team members' knowledge and skills (Boerma, Tsyvinski, and Zimin, 2021; Freund, 2022; Herkenhoff et al., 2024). Closest to this paper, Pearce (2022) uses technology classifications and citations to study changes to the team knowledge production function over time. However, this literature has largely been limited to studying innovation value. This is due to a lack of models and data capable of disentangling individual con-

[4]Using the data kindly provided online by Arts, Hou, and Gomez, 2021. They provide a dataset which identifies new words, and new n-grams in patents and which of these are later re-used. This allows them to capture both novelty and impact.

tributions to knowledge components. This chapter makes use of the model of team work developed in Chapter 1 which overcomes these limitations by utilising patent texts.

This paper joins a growing and important literature that looks to describe the innovation landscape, and how it develops over time using topic models. I extend the concept of building a map of innovation, developed in Fleming and Sorenson (2004), to the inventor, team and patent level. I do so by making use of high-dimensional patent text data. Carvalho, Draca, and Kuhlen (2021) study how firms and inventors either explore for new technologies or exploit existing ones, using a model of Latent Dirichlet Allocation to describe a firm or inventors position in the knowledge space. In a similar vein, Teodoridis, Lu, and Furman (2022) develop a Hierarchical Dirichlet Process at the patent level to map the knowledge space over time. Both papers can aggregate from the patent level to the inventor or firm unit, however they do not model within team dynamics. In contrast, I model collaboration directly through an LDA to disentangle inventor contributions and model the key producers of innovation, the teams themselves. This paper contributes to this literature by allowing for a more accurate representation of team production, by removing the assumption that inventors contribute uniformly within a team. This allows me to better characterise the role of each inventor. Therefore locating the team within the knowledge space to better characterise the development stage of the research fields on which the team is building.

Finally, I contribute to the literature on use of natural language processing models to capture breakthrough science and innovations. Arts, Hou, and Gomez (2021) and Arts, Melluso, and Veugelers (2025) developed the literature beyond using citation histories. They do so by identifying the new words created by patents and papers in order to measure novelty. They then capture which of these are re-used by future innovations to measure impact. Kelly et al. (2021) develop a method of identifying breakthroughs by comparing the similarity of patent texts to patents which came before and after. The concept of breakthrough in this paper builds directly on their foundation. The key contribution of this paper is to extend this to the team level to connect the novelty and impact of their innovations to the development stages of their research fields. This paper employs a two-stage approach to back out the required latent variables from text. There is a recent literature on inference concerns when using two-stage methods (Bandiera et al., 2020; Battaglia et al., 2024). However, as discussed

in the original paper, patent texts are highly dimensional and these concerns are reduced in this context.

Paper Outline The rest of the paper is structured as follows: Section 2 defines the theoretical framework; Section 3 outlines the process of inferring the knowledge space from text; Section 4 describes the empirical reduced-form strategy; Section 5 provides descriptive statistics and validation tests; Section 6 presents the main results; and Section 7 concludes.

2.2 A Framework for Team Innovation

Define $\mathcal{K}$ as a set of K knowledge classes.[5] Each class represents a specialised area of understanding. Inventors innovate by combining their knowledge on these classes. I model the innovation and writing of a patent as a single, unified process. There is a fixed vocabulary of words which inventors can use, denoted by V. Inventors use different words when describing different knowledge classes. This is captured by the probability distribution β_k for topic k across the vocabulary. β_{kv} captures the probability of using word $v \in V$ when discussing class k.

A 3-dimensional example is given by

$$\mathcal{K} = \{\text{Computing, Transport, Medicine}\}.$$

The words *hospital, doctor* and *syringe* are more likely to be used when describing a medical innovation than one about transport. One patent though may combine multiple classes. For instance, a drone to deliver prescriptions will likely use words correlated with both the medical and transport classes.

Denote $\Delta(\mathcal{K})$ as the knowledge space which is defined as the $(K-1)$ probability simplex over the set $\mathcal{K}$. θ is a point in the simplex, such that it represents a combination of knowledge classes. Let I be the set of all inventors. Each inventor is characterised by their knowledge profile θ_i. Formally, this is drawn from the

[5] No two knowledge classes are more similar to each other. This is a simplification that can be addressed with more complex models that allow for correlation between knowledge classes. Consult Blei and Lafferty, 2005 for further details.

knowledge space $\Delta(\mathcal{K})$ according to a Dirichlet distribution

$$\theta_i \sim Dir_{\Delta(\mathcal{K})}(\alpha),$$

where $\alpha \in \mathbb{R}^K$ is the non-symmetric Dirichlet prior such that $\alpha_k \neq \alpha_j > 0$. The support for a Dirichlet distribution is the set of K-dimensional vectors $\boldsymbol{x}$ where each $x_k \in [0, 1]$ and $\sum_{k=1}^{K} x_k = 1$. The value of the Dirichlet distribution is that each element in the support of a Dirichlet distribution can be treated as a K-dimensional discrete probability distribution.[6]

If the average α_k is low then the mass of the Dirichlet distribution lies in the corners of $\Delta(\mathcal{K})$. This means that inventors are more likely to hold knowledge on a few classes as opposed to being spread over many. In other words, inventors are more likely to be specialists than generalists as the average α_k tends to zero.[7] I allow for a non-symmetric Bayesian prior, so that on aggregate, certain knowledge classes will be more common.

A team $\tau \subseteq I$ is a set of m inventors who produce patent p together. When a team τ collaborates, they first choose the share of the workload to be performed by each team member. These shares are not constrained to be uniform across team members and some may contribute more than others.[8] I model this as a random draw where the team chooses a vector ω_p such that $\sum_{i \in \tau} \omega_{ip} = 1$ and $\omega_{ip} \geq 0$. Each ω_p is drawn uniformly at random. This can be modelled as a draw from another Dirichlet. This time with a uniform prior $\alpha = \mathbf{1}$. Drawn from the set of all possible workload divisions for m team members, denoted as Δ^{m-1}

$$\omega_p \sim Dir_{\Delta^{m-1}}(\mathbf{1}).$$

The team then produces a patent according to the following stochastic process.[9] The team first draws the number of words in the patent $N_p \sim G(\cdot)$.[10] Then for

[6] In fact the Dirichlet is the conjugate prior for the multinomial distribution, a feature that is utilised in defining the estimation method.

[7] This matches the literature by modelling inventors as more likely to be specialists than generalists.

[8] Inventors are often modelled as agents with a high level of autonomy over project choice and team participation (Akcigit et al., 2018) and allowing for these weights to be chosen optimally is an important next step.

[9] This process is outlined in greater detail in Chapter 1.

[10] This distribution G is irrelevant for the model. An appropriate approximation can be learnt from the observed set of patent lengths. Potentially this could be interesting over time since patents have become significantly longer throughout the period studied.

each word $n_{ip} = 1, \ldots, N_p$ the team draws an inventor $i \in \tau \sim \omega_p$ and from that inventor's knowledge distribution draws a class $k \in \mathcal{K} \sim \theta_i$. Given the corresponding knowledge class to word distribution, the inventor draws a word $v_{ip} \in V \sim \beta_k$. Each word in the patent is paired with a knowledge class, which produces a patent knowledge class distribution. Since the number of words in a patent is large, in expectation we can define the expected patent knowledge distribution. I denote the expected patent distribution as θ_p^e to simplify notation throughout the paper:

$$\theta_p^e \equiv \mathbb{E}[\theta_p | \tau, \omega_p] = \sum_{i \in \tau} \omega_{ip} \theta_i. \tag{2.1}$$

Therefore, synonymously to inventors, a patent can either be on a very specific topic, or a combination of many. Importantly, inventors, teams and patents now belong to one consistent space. This enables the counting of how much innovation exists in each local knowledge field.

The knowledge contained in the patent is a function of the inventors who produced it. However, given the stochastic process, the final patent distribution will not equal its expectation: $\theta_p \neq \theta_p^e$. Though it will likely be very close, since the probability that a given team τ produces a patent distribution θ_p is decreasing in

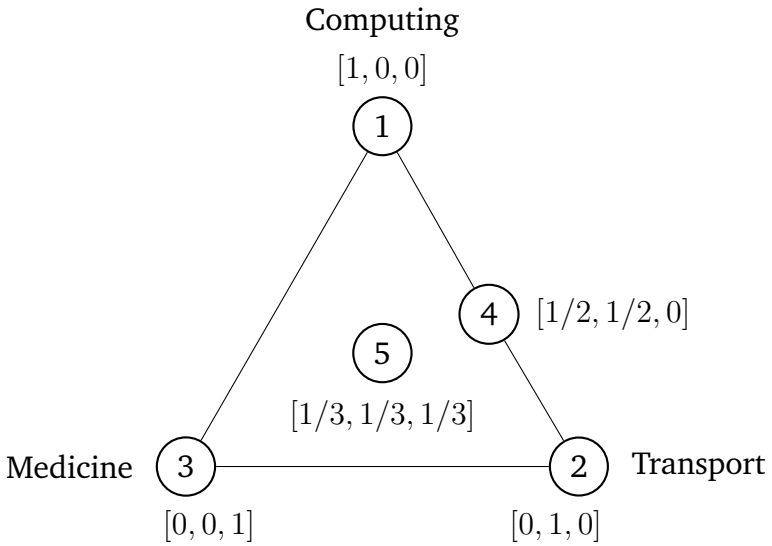
$$d(\theta_p^e, \theta_p) = ||\theta_p^e - \theta_p||. \tag{2.2}$$

The team first assigns roles within the team, which given the knowledge profile of each team member defines the expected outcome of their collaboration. The stochastic process by which the team generates the innovation is consistent with the idea of them pursuing a method of trial and error, in which each inventor tries many ideas and the probability of success is equal to their contribution weight.

Given the previous example of $K = 3$, the knowledge space is a 2-dimensional equilateral triangle and can be represented as in Figure 2.1. Each of the corners represent perfectly specialised profiles. An inventor or patent may split their knowledge over two of the classes, and hold no knowledge on the third, as in point 4. Point 5 represents the centroid of the simplex, and is a perfect generalist, sharing their knowledge equally over all classes.

FIGURE 2.1

THE KNOWLEDGE SPACE



Notes: Example of a 2 dimensional knowledge space over 3 knowledge classes. Each point 1-5 represents either an inventor or patent knowledge profile, since both are characterised in the same space. In the full model I use K = 50 classes. This example is informative as can be plotted in 2-D, and while the number of classes is small, there number of combinations remains infinite.

If inventors 1 and 2 were to collaborate and contribute equally such that $\omega_{11} = \omega_{21} = 1/2$, then in expectation they will produce θ_p^e at point 4 in Figure 2.1. Then given the random innovation process, all patents along the line between points 1 and 2 are feasible outcomes, however decreasingly likely as the distance from point 4 increases.

Within this space I define a local knowledge field for both teams and patents. I define a local research field for each patent as a closed ball of radius r centred at point θ given by[11]

$$B(\theta, r) = \{\theta' \in \Delta(\mathcal{K}) \mid \|\theta' - \theta\| \leq r\}. \tag{2.3}$$

This field is fixed over time, however the number of other realised patents belonging to the local research field can vary over time.

I define $\tilde{S}(\tau)$ as the team span: the set of all linear combinations of the team

[11]The choice of r is important here in the sense that r can determine which patents are classified as breakthroughs. I find no empirical difference in the regression results from changing r. It is also linked to the choice of dimension K. If you increase K, then if you want to keep the dimension of what is a breakthrough constant, r should be adjusted downwards.

members' knowledge distributions. Given the assumption that the weights ω_p are drawn from a uniform distribution, the team is equally likely to draw any patent in this set as their expected output, such that $\theta_p^e \in \tilde{S}(\tau)$. Formally I define the team span as the convex hull across team member distributions:

$$\tilde{S}(\tau) = \left\{ \sum_{i \in \tau} \omega_{ip} \theta_i : \sum_{i \in \tau} \omega_{ip} = 1, \omega_{ip} \geq 0 \right\}. \tag{2.4}$$

To define the local knowledge field for a team consider the Minkowski sum of $\tilde{S}(\tau)$ and $B(\theta, r)$. The resulting set is analogous to the local knowledge field at the patent level. In fact the local knowledge field for a team of one is defined identically. This sum expands the team span into the full K-dimensions of the knowledge space. The team knowledge field is in fact the full set of patent research fields in which they could patent in expectation.

$$S(\tau) = \tilde{S}(\tau) \oplus B(\theta, r) = \{x + y \mid x \in \tilde{S}(\tau), y \in B(x, r)\}. \tag{2.5}$$

Continuing with the example outlined previously, Figure 2.2 demonstrates how inventors, teams patents and research fields lie in one consistent space. Panel (A) shows an example of a patent's local research field. The plot is fixed at the year patent p (shown in black) was published and there were five examples of prior work in the local research field. Panel (B) shows an example team of three members, the interior shaded area represents their span $\tilde{S}(\tau)$. Each inventor lies in one of the vertices of the interior shaded triangle. The outer perimeter defines their local knowledge field $S(\tau)$.

2.2.1 Characterising Patent and Team Fields

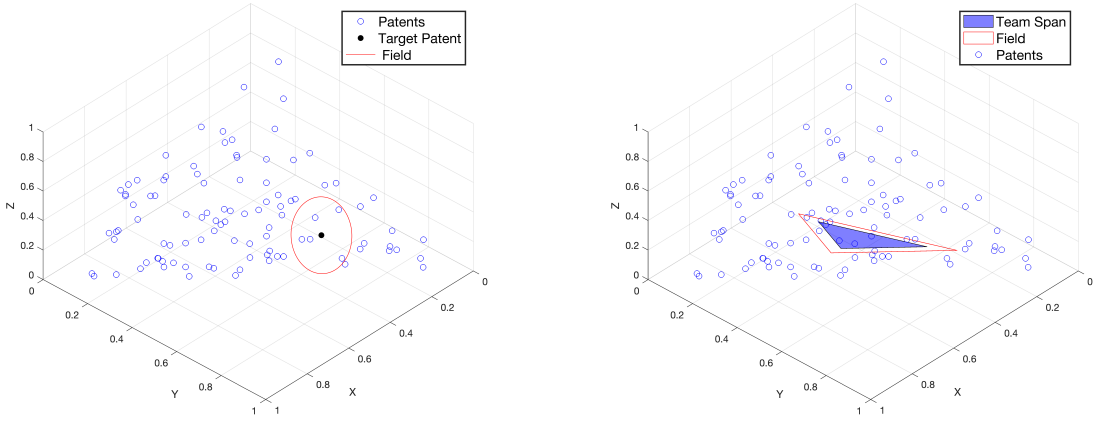
This method backs out a latent representation of both a patent's research field, and a team's knowledge field: the set of all patent research fields on which they work in expectation. Once learnt from data, the econometrician can apply any function they desire to these objects in order to describe them and explain their role in innovation.

Define the set P_t as the set of all patents published in the global knowledge space up to and including period t. Define the following count for the number of these patents which belong to the local research field of patent p at θ_p.[12]

[12] A detailed explanation of how I count these objects empirically is provided in Appendix B.4.

FIGURE 2.2
A LOCAL KNOWLEDGE SPACE

A) PATENTS B) TEAMS



Notes: The example patents and inventors are generated from a Dirichlet distribution with $\boldsymbol{\alpha} = [2, 1.5, 1]$*, which leads to the distribution across the knowledge space being weighted towards the bottom-left corner. The left panel shows the research field for a target patent, shaded in black. The right panel shows the knowledge field for a team of three inventors.*

$$n_{pt} = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in B(\theta_p; r)\Big) \tag{2.6}$$

$\mathbb{1}$ denotes an indicator function which is equal to one when the condition in parentheses is met. I propose the following breakthrough measure at the patent level, which is an adjusted percentage change to allow for zero patents either before, after or both.[13]

$$\text{breakthrough}_{t(p)s} = \left(\frac{\text{post-count}_{ps}}{1 + \text{prior-count}_{pt} + \text{post-count}_{ps}} \right) \tag{2.7}$$

For a patent produced in t, prior-count$_p$ aggregates each n_{ps} for $s \leq t$ and post-count for all patents produced in $s > t$. Holding prior-count$_p$ constant, the breakthrough score of a given patent p is increasing in the number of patents which came afterwards. It increases non-linearly, with decreasing returns, such that early entrants contribute more than late comers. Figure B.1 gives an ex-

In short, I first slice the data by the maximum distance within the team field. I then check for the remaining patents which belong to the team field by checking the distance from each patent θ_q and the exterior of the team field.

[13] In section 2.5, I compare the patents identified as breakthroughs by equation 2.7 to the literature to demonstrate the precision of this method.

ample that also demonstrates that the curve of the breakthrough measure with respect to post-count$_p$ flattens as the prior-count$_p$ increases.

I classify breakthrough patents as those that land in the top quartile of residuals from a regression of breakthrough$_{t(p)s}$ on a set of application year dummies. I use these residualised values for the following reason. The measure presented in equation 2.7 is the raw breakthrough measure, however as made clear in Hall, Trajtenberg, and Jaffe (2001), when working with patent outcomes it is important to control for the fact that they are right-coded in time. Patents produced recently have not had enough time to be revealed as breakthroughs, since the patents that build on them have not yet arrived.

Patents produced in areas with few pre-existing works are novel, but only those which post-publication see a significant increase in the number of patents belonging to their local knowledge field are breakthroughs. This is similar in concept to the breakthrough measure proposed by Kelly et al. (2021), however uses a spatial dimension that is easier to track and visualise over time. One key contribution of this paper is to incorporate the teams who produce these patents.

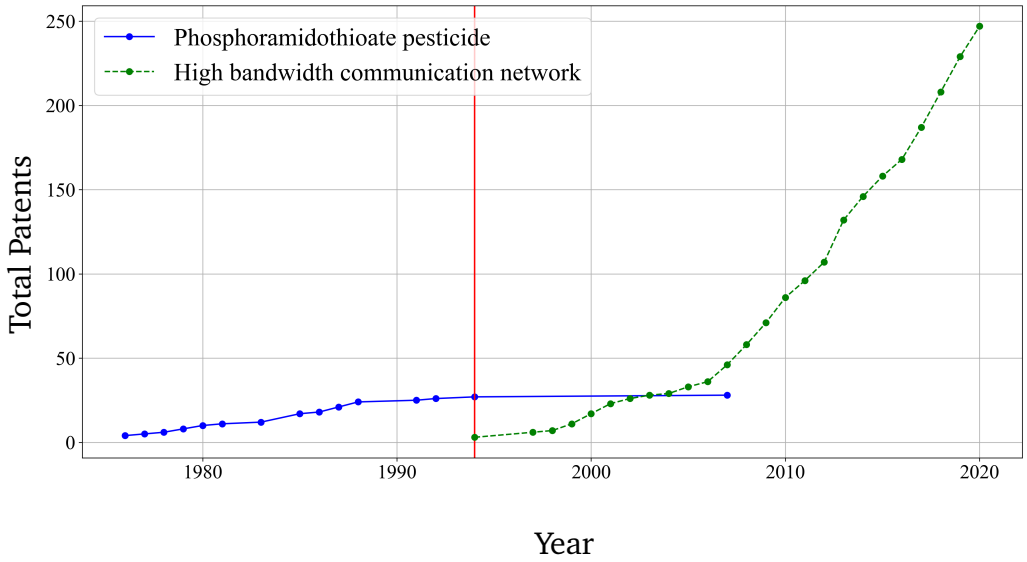
Figure 2.3 provides two examples of patents, one classified as a breakthrough and the other not. Both patents were applied for in 1994, however in different locations in the knowledge space. Following their publication the two research fields developed along very different paths over time. The Y-axis plots the total number of patents within each patent's local research field. Patent *US5612948* titled *High bandwidth communication network* scores very highly. After their publication, their area of the knowledge space grew into a vibrant research area. Whereas the patent *US5597812* titled *Phosphoramidothioate and process of use to combat pests* developed on an area showing slow growth, and on which only one future patent develops.[14]

At the team level you can define the synonymous count. The team knowledge field represents the set of patent research fields in which the team could patent in expectation. The sum in equation 2.6 essentially counts the quantity of unique prior work which exists in each of these knowledge fields.[15]

[14]On further examination, this type of pesticide was widely used in the past to control various pests. However, due to their high toxicity and potential environmental risks, their usage has significantly declined.

[15]The measure counts the unique count in that it does not double count for overlapping fields.

FIGURE 2.3
EVOLUTION OF LOCAL KNOWLEDGE FIELDS



Notes: Two cases of low and high breakthrough patents, using the estimated knowledge space. This is the raw count, and does not remove year fixed effects. The vertical red line identifies the publication year of both patents. The Y-axis records the total number of patents in each target patent's research field.

$$n_{\tau t} = \sum_{q \in P_t} \mathbb{1}\Big(\theta_q \in S(\tau; r)\Big) \tag{2.8}$$

I approximate the volume of a team local knowledge field using the following equation[16]

$$\text{Volume}\Big(S(\tau)\Big) = v_\tau = \sqrt{m} \times (\mu \cdot D_{\text{max}} + (1-\mu) \cdot D_{\text{mean}})) \tag{2.9}$$

Where m denotes team size and $\mu \in [0, 1]$.[17] Here D_{max} is the maximum distance between any two team member distributions, and D_{mean} is the average across all pairwise combinations of team members. This is not a function of r since that is constant across teams.

[16]Measuring the volume in high dimensional space is challenging, and there are alternative ways to do this. For example, using a set of uniformly distributed points and sampling via MCMC.

[17]In the final model I set μ equal to 0.7, to emphasise the total breadth within the team span. However, I have run the model for many alternative levels and the results don't change.

2.2.2 Testable Predictions

Using this framework I derive a set of testable hypotheses. In appendix section B.3, I present a more rigorous derivation of these hypotheses which arise directly from the innovation production process.

As a research field develops, it moves up the ladder of development, and this represents an increase in the quality of existing knowledge. The cost of producing follow-on innovations is then decreased as inventors build endogenously on the work that came before them (Grossman and Helpman, 1991). However, faced with an increasing knowledge stock, inventors suffer from the "burden of knowledge". As the innovation frontier expands, this represents an increased cost to inventors of reaching and developing on this frontier (Jones, 2009). As a team's local knowledge field populates with patents, it becomes harder to produce a truly innovative idea (Bloom et al., 2020). The compound definition of a breakthrough as both novel and impactful leads me to the following hypothesis at the patent research field level.

Hypothesis 1. *There is an inverted-U shaped relationship between the quantity of prior work in a patent's research field and the likelihood of the patent being a breakthrough.*

This implies that as the number of existing patents within a research field increases, the probability that a new patent is a breakthrough first increases due to knowledge accumulation, then decreases after a certain point due to saturation.

In that case, how can a team produce a new breakthrough idea? For a team composed of inventors covering a well-established research area, finding a novel idea is challenging. Increasing the novelty of their work might require removing a member contributing knowledge from the most developed fields. This hypothesis may seem contradictory with the well established idea of increasing the number of potential combinations between inventors. The idea is that the set of combinations which the team can create is determined by the knowledge profile of each team member. If one team member comes from a well established field, their presence in the team may drag the team closer towards this established fields. By removing certain members you may reduce the set of combinations, however in doing so free the team from more established paradigms and enable them to be more novel. On the other hand, for a team in an under-explored area, this same adjustment would be harmful. It would strip away the limited knowl-

edge they possess, setting them further back on the ladder of development.

I propose to test this with an additional hypothesis at the team level. Importantly, when teams remove (or add) a member they also change the size of the team's field, in doing so they change the quantity of potential combinations. The effect of building on a few patents dispersed across a vast set of possible combinations is likely different from building on the same number within a smaller, more concentrated field. The following hypothesis includes this feature by using variation in the density of prior work within a team's field.

Hypothesis 2. *The impact of reducing the density of patents in a team's knowledge field on the probability of their next patent being a breakthrough depends on the development stage of their initial field:*

- *If the team spans an advanced research area, moving to areas with a lower density of prior work increases their breakthrough probability.*
- *If the team spans an early-stage research area, moving to areas with a lower density of prior work decreases their breakthrough probability.*

The rationale for why is as follows. Prior work enhances the impact of an innovation; thus, when a team incorporates more related work, the quality of their innovations improves. However, for a team to produce a truly innovative idea, the existence of prior work in the same field is a barrier. Locally to the patent, this is intuitive since it is now not the first to market. At the team level however this result is more subtle. By reducing the density of prior work within their local knowledge field, the team will draw ideas from less populated areas of the knowledge space. The idea being that by reducing the presence of prior work, the team is freed from established paradigms, and are capable of producing a breakthrough idea.

2.3 Inferring the Knowledge Space

I first outline the data and sample over which the model is approximated. I then introduce the Bayesian model of Natural Language Processing used to infer the knowledge space. This allows me to count the quantity of prior work within a team's local knowledge field as to test both hypotheses.

2.3.1 Data and Sample

I build the knowledge space from US patent data from *patentsview*, the online data base for the United States Patent and Trademark Office (USPTO). I restrict the sample to teams who applied for their first patent after 1990, and their last prior to 2011. I build the sample around three types of teams, which I combine into a panel of team, patent observations.

The first team type are those teams which are treated by the premature death of a co-inventor. The premature death of an inventor is determined using the dataset provided by Kaltenberg, Jaffe, and Lachman (2021). I define a premature death using the following logic. I take one unique death date per inventor[18], and classify premature as an inventor who dies within three years of patenting with the team. This defines a treated inventor, and treated team. I then search for teams which return to patent, within five years, for two cases: they return minus the deceased inventor, or having replaced that inventor with one other. Teams which return with two or more new inventors are dropped from the sample. Given the delay in producing a patent, returning in less than five years is relatively fast to turn around a new patent. I claim that the death was a quasi-natural experiment in changing team composition, I discuss this strategy in more detail in section 2.4.2.

I add to this sample two additional types of teams which act as controls. The first are pure controls: a team which never adds or removes a member. This group of teams never appear again either without one or more members, or having added one or more new ones. The second are those that first patent with m inventors, then that after that team publishes their final patent the same inventors return, with one additional member, again within five years. The first set of baseline controls provide a baseline comparison for whether teams change their output dynamically. The second provide an endogenous team composition change that allows me to study adding new members as a robustness check. In total I find 353 teams treated by a premature death who return without the deceased inventor, 2200 treated teams that replace that inventor with one other. Then to find the controls I draw from a random sample of 300,000 teams according to the criteria above. I find 6400 baseline control teams and 980 teams which add one

[18]This data set was produce by scraping four well known US public record databases, for many inventors they scraped multiple potential birth and death dates. They score each one according to their belief that it is an accurate measure. I take the maximum observation with a maximum score. For more details see the original paper.

new member. However, since I am using a conditional logit model, I estimate the treatment effect on teams which switch outcomes at least once. In other words they produce least one breakthrough. The final sample includes the following split: 72 teams which don't replace the prematurely deceased inventor, 510 which do replace them and 1709 baseline control teams.

I extract the full patenting history of each member of every team. I train the LDA on this sample of 408,774 patents written by 270,065 inventors. This sample contains patents and inventors for which I don't track their entire history, however they help provide a precise measure of knowledge classes for the target sample. To measure on what fields do patents build I populate this space with a random draw from the universe of USPTO patents. I extract over 2.2 million USPTO patents, approximately one third of the universe of USPTO patents grants over the period studied.[19] I populate the knowledge space with this random sample by treating each patent as if it were a new author, who patented one solo paper. Then taking the trained model and learnt knowledge classes, I fit each patent into the estimated knowledge space.

I combine additional data for the robustness check, and additional sections demonstrating the knowledge space. Firstly whether they are a breakthrough or achieve a certain direction. Kelly et al. (2021) classify the universe of USPTO patents from 1976-2014 as whether they are a breakthrough, or not. I measure three innovation directions exogenously. They are three binary indicators for whether a given patent achieves that purpose, or not. The first is whether that patent is a labour saving technology (Mann and Püttmann, 2023). Secondly does that patent mitigate climate change which is measured as whether that patent is awarded the YO2 patent class (PatentsView, 2024 and finally does that patent target improving cancer diagnosis or treatment (Cancer Moonshot: USPTO, 2024).

2.3.2 Latent Dirichlet Allocation

Patent texts are increasingly used to describe the knowledge content of innovations, and the innovation literature has begun to borrow and develop models from the computer science literature in order to answer new questions on science and technology. Patent number US9939179 begins their detailed descrip-

[19]This is a rough calculation. To determine the denominator in this calculation I use the fact that there were 6,901,791 patent's granted between 1976 and 2020

tion with the following:

> *However, one of ordinary skill in the art will recognize that the invention is not necessarily limited to refrigeration systems. Embodiments of the invention may also find use in other systems where multiple compressors are used to supply a flow of compressed gas.*

This quote demonstrates that the patent texts are informative on the knowledge content beyond a simple title or CPC classification. The text describes features of the innovation that can be applied to other fields. In order to extract this information into a empirically feasible dimension I use a model of Latent Dirichlet Allocation (LDA). LDA models were first developed by Blei, Ng, and Jordan, 2003 and have become a popular method of NLP. Consider this a brief and intuitive overview of how an LDA infers a set of parameters which approximate the knowledge space. For a full description please refer back to Chapter 1 on "Modelling Collaboration Through Patent Texts" for a more complete description.

The model is built upon the paradigm of observing the set of patent texts, and proposing a hierarchical Bayesian model. This allows the researcher to infer a set of latent parameters which govern how that set of texts was produced. The model identifies many parameters jointly, most importantly: inventor and patent knowledge class distributions and each inventors' contribution weight to each patent.

Prior to estimating, I preprocess the text in order to improve the model inference, by stemming and removing stopwords Sarica and Luo (2020). The words contained in a patent describe its design and use. The LDA model reduces the dimension from over 250,000 words in the raw patent texts to infer a distribution for each knowledge class across the set of unique words. The logic here is that certain knowledge fields use specific words, jargon, more than others when describing objects or problems from their field. For example, someone describing a medical patent is more likely to use the words blood, cells and syringe than someone talking about vehicles, who is more likely to use car, wheel and door.

The model uses the knowledge classes as a dimension reduction technique since a distribution for all inventors across all words is harder to manage both conceptually and computationally. The words presented are stemmed as part of the text cleaning process, e.g. the word *imag* represents image, images and imaging. The model does not attach labels to the knowledge classes, though they can be approximated using GPT technologies which analyse the word weights.

I build on the *Gensim* python package (Mortensen, 2017) which trains the unsupervised machine learning model by implementing a method of Variational Bayes. The objective is to infer from patenting histories which team member was most likely to have contributed each word and with which knowledge class. In doing so, infer the inventor knowledge distributions and their contribution shares to patents. An inventor with a long history of producing transport patents will be more likely to have contributed the words vehicle, destination and route. If a given patent includes many words highly correlated with the transport class, the model will give a larger contribution share to that inventor.

Table 2.1 provides the hyper-parameters which govern the estimation process.

TABLE 2.1

LDA PARAMETERS

K	η	Iterations	Passes	γ
50	1/K	350	100	0.001

Notes: K is the number of knowledge classes. η the Bayesian Dirichlet prior on the knowledge class to word distribution. Iterations sets the number of cycles used to update the knowledge class distributions, passes are full the number of times the model goes over the entire dataset, and the gamma threshold sets the stopping point when the difference between topic updates is sufficiently small. The model has been run various times changing these parameters, and the results remain similar. Both η and γ are set to the `Gensim` default values. For more details consult the ATM package documentation online.

A key parameter of choice is the number of knowledge classes. I choose here $K = 50$, which is close to the optimal number of topics chosen by Teodoridis, Lu, and Furman (2022). They back out an optimal number of 79 classes.[20] η is the prior for the knowledge class to word distribution and is assumed to be symmetric. I allow the model to back out the Bayesian prior α, which I assume is asymmetrical.

The perplexity measure is the standard measure used within the topic modelling literature to evaluate the quality of topics estimated. The perplexity score measures how well the model predicts the words in the documents based on the learned topic distributions. In other words, how well the model captures the underlying structure of a set of documents. A lower perplexity score indicates that the model has a better ability to generalise to unseen data, and convergence

[20] As discussed in Chapter 1, since an inventor's knowledge profile and contribution share are both continuous, bounded variables, in theory, the choice of K is not a key determinant of the later empirical analysis.

indicates that the LDA has effectively learned the topic structure of the patents.

Figure B.4 plots the estimated Bayesian prior over the knowledge classes and the 5 words with the largest weight within the distribution for that class. We see variation across classes, which allows for some classes to be over-represented, which will reflect aggregate innovation direction across the time period.

2.4 Empirical Strategy

I present a set of regression models to test both hypotheses derived in section 2.2. To tackle the research question on how teams build on prior work I first start at the patent level. I test the relationship between the quantity of prior work on which a patent develops and the probability that patent produces a breakthrough.

2.4.1 Hypothesis 1: Patent Level

Here the dependent variable varies at the patent level, where each patent maps into one team τ and application year t. The regression is run as as a standard logit model to predict whether patent p from team τ in year t is a breakthrough, or not. The full specification is given by

$$Pr\Big(Y_{\tau t(p)} = 1 \mid X'_{\tau t(p)}\boldsymbol{\psi}\Big) = \frac{\exp\Big(X'_{\tau t(p)}\boldsymbol{\psi}\Big)}{1 + \exp\Big(X'_{\tau t(p)}\boldsymbol{\psi}\Big)} \tag{2.10}$$

where

$$X'_{\tau t(p)}\boldsymbol{\psi} = \beta_0 + \delta_t + \beta_1 n_{pt} + \beta_1 n_{pt}^2 + \beta_2 d_p + \beta_3 m_\tau \tag{2.11}$$

The main parameters of interest are β_1 and β_2. $\beta_1 > 0$ and $\beta_2 < 0$ are consistent with an inverted-U shape. The model controls for the randomness in innovation by including the distance between the realised patent distribution and the expected value in $d_p = d(\theta_p^e, \theta_p)$ as defined in equation 2.2. I include the team size as a control with m_τ. I include year fixed effects for multiple reasons. They control for the fact that breakthroughs are right-coded in time: patents published recently have not yet had chance to be realised as breakthroughs.

2.4.2 Identification Strategy for Team Outcomes

The headline result is how team innovation outcomes change after moving into a new area of the knowledge space and therefore building on a different set of prior work. I utilise two types of changes to identify the effect of shifting the location of a team. Both follow the premature death of a team member. I define premature as having died within three years of patenting. The number three is chosen as in the USPTO raw data, teams on average patent every three years. Therefore if an inventor dies within three, it is reasonable to assume that on average this would change their next patent outcome. This definition is therefore based around them being active, not their age or health status, and is in line with the literature Azoulay, Fons-Rosen, and Graff Zivin, 2019. Denote the initial team, prior to the premature death as τ_1. This team must return to patent within 5 years denoted as τ_2. This is to define a cap on the number of years in which they must return. The identifier τ is now a unique id for each pair (τ_1, τ_2). Either τ_2 consists of the original team minus the deceased inventor ($\tau_2 = \tau_1 \setminus \{i\}$), or they replace i with one other inventor j ($\tau_2 = (\tau_1 \setminus \{i\}) \cup \{j\}$). Therefore I only allow for small changes to team membership.

I define the measure $D_{\tau t} = n_{\tau_1 t} - n_{\tau_2 t}$ to measure the change in the quantity of prior work on which the team is building, following their shift in the knowledge space.[21] The identifying assumption here is that the death is an unexpected event, where the impact on the team's knowledge field is captured by $D_{\tau t}$. Notice that the treatment is time dependent. If a team member prematurely dies in period t then $D_{\tau t}$ measures the contemporaneous change in the quantity of prior on which the team builds. However for all future periods this variable captures the knowledge foregone by the untimely death. The idea being that if the inventor had not passed away, the team could have continued to patent in those research fields. I control in the regression for the first team's count $n_{\tau_1 t}$, such that β_1 captures the effect of removing existing patents from the team span, conditional on the prior quantity.

2.4.3 Hypothesis 2: Team Level

This model is run on a team patent panel. Each patent is a new period s, such that the team τ is repeated over their 1^{st}, 2^{nd}, 3^{rd} patents and so on. I then

[21] To ensure the logit model converges, I winsorise the top 1% of both the total and direction counts, $n_{\tau t}$ and $n_{\tau t}(z)$ respectively. This caps the maximum value at the 99% value, to reduce the effect of outliers.

predict the probability that $Y_{\tau s}$, a team's nth patent, is a breakthrough.

$$Pr\Big(Y_{\tau s} = 1 \mid X'_{\tau s}\boldsymbol{\psi}\Big) = \frac{\exp\Big(X'_{\tau s}\boldsymbol{\psi}\Big)}{1 + \exp\Big(X'_{\tau s}\boldsymbol{\psi}\Big)} \tag{2.12}$$

The full set of independent variables is given by,

$$X'_{\tau s}\boldsymbol{\psi} = \alpha_\tau + \mu_s + \delta_t + \beta_1 n_{\tau_1 s} + \beta_2 D_{\tau s} + \beta_3 v_\tau + Z'_p\boldsymbol{\delta} \tag{2.13}$$

Z'_p includes a set of controls for each patent, which vary at either the team or firm level. I control for the distance between the expected patent distribution and realised outcome, as described for equation 2.11. I then introduce a set of controls that help me claim conditional independence of the treatment. The death of a co-inventor and their replacement (or not) will change other features of team composition which may determine patent outcomes. To control for a set these (although not exhaustive) I control for the following. The gender ratio of team members, the average experience level[22], the square of average experience, the race diversity within the team[23]. In addition, I control for the rolling three-year average number of inventors employed at the institution to which the patent is awarded. This is a relevant control since large firms and universities may have different outcomes, for example due to capital resources, but they also have easier access to replacement inventors in case of a premature death.

Hypothesis 2 requires that the density of patents changes within the team local knowledge field. Therefore I introduce a control for the volume denoted v_τ, as defined in equation 2.9. $n_{\tau_1 s}$ controls for the quantity of prior work within the team field of the initial team, prior to the inventors premature death. β_2 then captures the treatment effect of removing patents from the team's knowledge field. In other words, the effect of moving them into a less explored area of the knowledge space.

To test the compound hypothesis, I split the sample of teams into quartiles of prior work $n_{\tau_1 s}$. I then run the same regression as specified in equation 2.12 for each quartile separately. For those teams initially building on a lot of prior work ($n_{\tau_1 s}$ high), $\beta_2 > 0$ is consistent with the gain from them moving into

[22]Experience is measured by the number of patents each inventor has collaborated on prior to the patent in question

[23]Race diversity is measured as the Shannon index $H = -\sum p_i \log p_i$, where p_i is the proportion of inventors belonging to each race.

under-explored areas and drawing more novel ideas. Conversely, for those teams initially building on a little prior work ($n_{\tau_1 s}$ low), $\beta_2 < 0$ is consistent with them losing out by having fewer prior examples to incorporate, reducing the impact of their work.

TABLE 2.2

DESCRIPTIVE STATISTICS

LDA sample				
Patents	Obs	Mean	Min	Max
Team size	408774	3.423	1	76
% Breakthrough	408774	0.260	0	1
Specialisation	408774	0.455	0.150	0.975
Concentration	408774	0.047	0	0.928
Inventors	Obs	Mean	Min	Max
Teams	270,065	3.078	1	767
Patents	270,065	5.181	1	4549
Specialisation	270,065	0.533	0.028	1
Contribution weight	270,065	0.263	0*	1
Treatment sample		1990-2010		
	No Replace	Replace	Control	
Treatment Status	60	447	1514	
	Obs	Mean	Min	Max
Team size	9,498	2.597	1	20
Team patents	9,498	7.239	2	51
% Breakthrough	9,498	0.45	0	1
Total Count	9,498	196.856	0	3321
Volume	9,498	0.683	0	4.69
Density	7,336	323.937	0	35327.19

Notes: Volume is defined by the square root of team size, multiplied by the weighted average of the maximum and mean distance between team member knowledge profiles. Total count is defined as the number of patents within a team's or patent's knowledge field. Density is defined as total count divided by the volume. * since this is approximately zero in the data. The treatment sample split is conditional on them being part of the final conditional logit sample- they have at least one breakthrough patent and non-missing values for the controls.

2.5 Describing the Knowledge Space

In this section I present a set of new descriptive statistics which are feasible in the knowledge space and provide important insights into team innovation. I also use this as a chance to validate the space by comparing the results to data taken from the literature. Table 2.2 shows two sets of descriptive statistics. The first panel describes the sample used to train the LDA model. The second panel is a sub-sample of the first, and describes the teams and patents used in the main reduced-form regression model.

We see that the team size is on average one person fewer in the treatment sample. Even though many teams replace their inventor, some do not, and this in part reflects that fact. In addition the percentage of breakthroughs increases, which reflects that the conditional logit model is identified only for the switchers, so teams that produce at least one breakthrough.

2.5.1 Aggregate Statistics

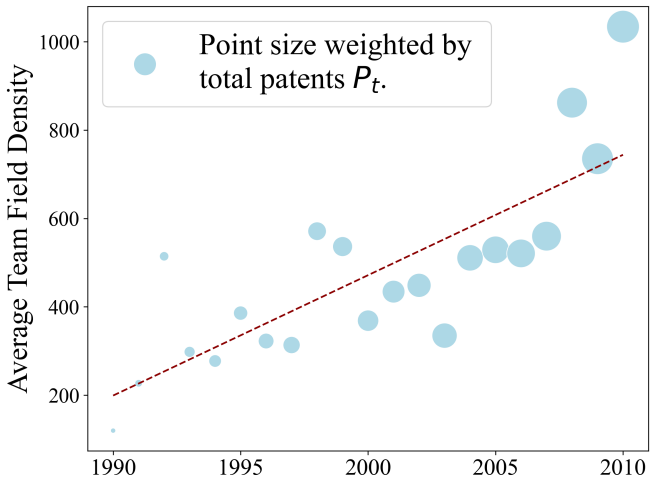
This paper examines how the maturity of a research field determines innovation outcomes for teams working in that area. While the knowledge profile and team span is constant over time, innovation arrives dynamically to the knowledge space. Therefore the research area on which a team works develops over time.

According to Bloom et al. (2020), innovative ideas are getting harder to find. This can in part be explained by an increasingly populated knowledge space. Figure 2.4 plots the average density of a team's local knowledge field across the sample, for teams which patent for the first time in each year. The density of prior work within a team's local knowledge field is defined as $n_{\tau t}/v_{\tau}$.

As the innovation frontier expands, the number of patents within the knowledge space increases. This does not however mean that the number of patents within a team's local knowledge fields increases mechanically.[24] Teams may endogenously respond and locate themselves in less populated areas. I show that for teams producing their first patent, the density of prior work within their field has increased, on average, over the sample period.

[24] In order for the result to not be mechanical, it does however assume that there is variation in the distribution of prior work across the knowledge space. Or that there is sufficient space for teams to locate themselves on top of a given quantity of prior work. Given that the space is continuous, I argue that this assumption is reasonable.

FIGURE 2.4
AVERAGE KNOWLEDGE FIELD DENSITY



Notes: Density is defined at the team level, as the number of patents within their local knowledge field normalised by their volume. Volume is defined by the square root of team size, multiplied by the weighted average of the maximum and mean distance between team member knowledge profiles, as given in equation 2.9. I then find the average density for each year, of team's which patented for the first time in that year. The size of each marker is weighted by the total number of patents in the knowledge space in each year.

This result can be examined further by looking at each part of the volume equation given in equation 2.9 and the each part of the density ratio. The number of patents within new team's local fields is increasing over time, while the volume of team fields is relatively constant. This is an interesting result, given that team size is increasing. For the volume to remain constant then, teams must be combining inventors who are closer together, such that the maximum and mean distance between members is decreasing. I show that this is in fact the case and the comparison across team statistics and the breakdown of the volume measure can be seen in Figure B.6.

2.5.2 Breakthrough Patents

This paper presents a novel empirical concept for breakthrough research fields. Table 2.3 provides a set of validation statistics to demonstrate the empirical power of the framework.

This paper develops on the work in Kelly et al. (2021) and using their data I find the correlation between their binary breakthrough classification and the one produced in this paper. I find a positive correlation of 0.234. A positive correlation

TABLE 2.3
VALIDATION OF BREAKTHROUGH PATENTS

Kelly et al. (2021)	Correlation between breakthrough classifications	0.234^{***}
	Corr. between pre-count$_p$ and breakthrough score	-0.121^{***}
	Corr. between post-count$_p$ and breakthrough score	0.226^{***}
Arts et al. (2021)	%Δ new re-used words in breakthrough patents	8.67^{***}
	%Δ new re-used bi-grams in breakthrough patents	47.6^{***}
	%Δ new re-used tri-grams in breakthrough patents	44.1^{***}
Citations	%Δ forward citations for + $\Delta 1\%$ in post count$_p$	$2.07\%^{***}$
	%Δ backward citations for + $\Delta 1\%$ in prior count$_p$	$1.09\%^{***}$
	$\Delta\%$ forward citations for breakthrough patents	$16.2\%^{***}$

Notes: Validation statistics using UPSTO citation data and existing patent novelty literature. The average number of new words, bi-grams and tri-grams used is 1.53, 5.85 and 8.08 respectively. The first panel displays the pairwise correlation coefficient. The second and third panels present log-log regression coefficients from a model which controls for application year and cluster dummies.

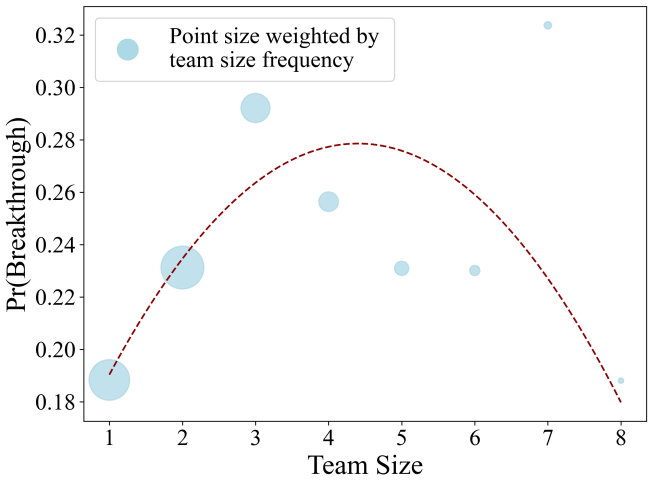
is expected, however the correlation is weakly positive. The correlation between the Arts, Hou, and Gomez (2021) and Kelly et al. (2021) is 0.28 for backward similarity and 0.29 for forward similarity. Therefore it is in a similar ball park for these two established measures, however further study is required to explain the differences in detail.

In addition, using the Arts, Hou, and Gomez (2021) data I first show that patents which I classify as breakthrough patents contribute 8.67% more new words which then go on to be re-used by future patents. Therefore these patents are relatively more novel, but they also have an impact by directing future research. They also introduce significantly more new combinations of existing words, 47.6% new word pairs, and 44.1% new-three word tuples. This result speaks to the central premise on how innovation occurs, through recombining existing knowledge.

Finally, I find that for each additional 1% of patents to enter the local knowledge field of a patent after its publication, the target patent receives 2.07% more citations. This elastic response points to the existence of knowledge spillovers between local patent sub-fields. This logic also holds for backwards citations where for each additional 1% of patents already present in a local knowledge field when a patent is produced, the target patent makes 1.09% more backward citations.

Having validated the breakthrough measure I replicate the first result from the

FIGURE 2.5
BREAKTHROUGH INNOVATIONS AND TEAM SIZE



Notes: The Y-axis plots the percentage of patents classified as breakthroughs, produced by teams of each discrete team size from 1-8. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Point size is weighted by the frequency of team sizes, since they are discrete bins and not equally sized.

literature in this unifying framework for teams. Wu, Wang, and Evans (2019) show that small teams disrupt science, while large teams develop it.[25] Figure 2.5 plots the probability of producing a breakthrough by team size. I plot up to a team size of 8 as this corresponds to 99% of the data. The graph confirms that teams outperform working alone, as all team sizes above 1 outperform solo patents. Teams of 3 work best, and team size is negatively correlated with breakthroughs beyond that.

2.5.3 Contribution Weights

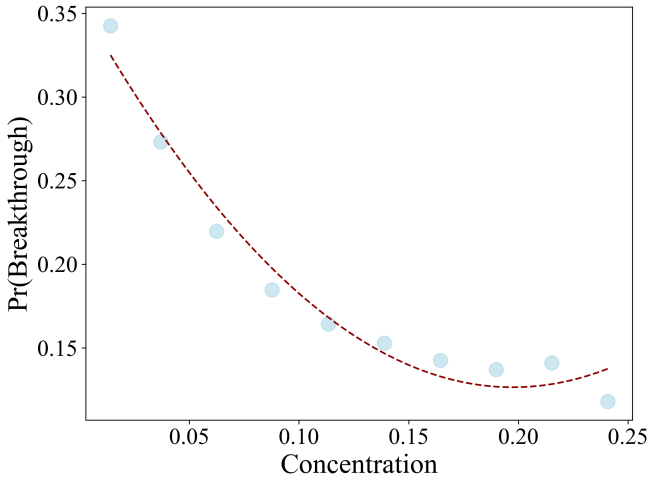
Table B.5 provides the results from a validation exercise on the contribution share inferred for each team member. The idea behind the validation test is the following. The patenting history for inventors for whom I back out a relatively large contribution share within the team, should be a stronger predictor of other patent characteristics. Using a random forest prediction model, I show that the technology class history for the lead inventor, compared to the second, is

[25]This paper uses a measure of innovation disruption. Disruption is measured by examining the citation patterns of future papers that reference a given paper. Specifically, they calculate a "disruption score" that reflects the extent to which a paper makes prior work obsolete or shifts the research direction.

a significantly stronger predictor of the technology class awarded to the target patent. This validation exercise is the same as that employed in Chapter 1, in section 1.5.

Having validated this measure I replicate a second well known result from the literature on team composition and breakthrough innovations. Xu, Wu, and Evans (2013) show that hierarchical teams produce fewer breakthroughs than teams in which members contribute more equally. Chapter 1 deals with this question in detail, and I have replicated this result again to show that this trend holds for this sample.[26] Figure 2.6 shows that teams which share contributions equally tend to produce more breakthroughs.

FIGURE 2.6
BREAKTHROUGH INNOVATIONS AND CONCENTRATION



Notes: Concentration is measured by taking the vector of contribution weights and finding the euclidean distance from a vector of length m (team size) in which all inventor contribute $1/m$. The Y-axis plots the average breakthrough value for 10 equally sized bins of the concentration measure. The breakthrough classification is based on Equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent).

2.6 Main Results

I first present a set of results that demonstrate how team innovations change in response to them pivoting to new research fields. This first sub-section can

[26] I demonstrate this result by taking the vector of contribution weights ω_p and finding the euclidean distance from the vector of length m in where all inventor contribute $1/m$. This measure is increasing in the concentration of the inventor contributions, and is minimised at 0 when all team members contribute equally. Consult Chapter 1 for more details

be skipped for those readers interested in the main breakthrough results. I then present the main results on how teams produce breakthrough innovations to test the hypotheses presented in section 2.2.

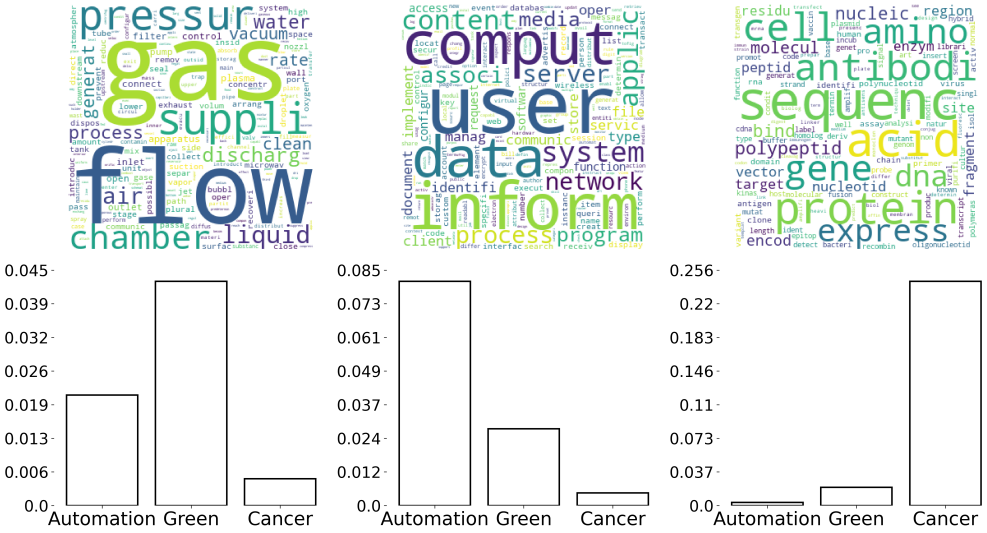
2.6.1 Knowledge Content of Team Innovations

In this section, I examine how the knowledge content of team innovations shifts as teams move into new research fields. I introduce three binary classifiers for the knowledge content of each innovation, which effectively segment the high-dimensional knowledge space into binary categories. This partitioning allows me to fit these classifiers into a regression model to show how team composition influences innovation outcomes. As team members are added or removed, the team's potential knowledge combinations change, and therefore the research fields available to them. This section illustrates how the knowledge content of their patents also depends on the history of prior work in each research field.

Each patent is classified by z_p where $z_p = 1$ if patent p achieves direction z. For this paper I take three exogenous classifications of whether each patent in the knowledge space achieves that purpose, or not. The directions are the following. Does the patent save labour? Does the patent improve cancer treatment? Does the patent mitigate the negative effects of climate change? These classifications are taken as exogenous (PatentsView, 2024; Mann and Püttmann, 2023; Cancer Moonshot: USPTO, 2024). Further details are presented in the data section 2.3.1. I combine the three directions in order to show how team innovations respond to past work, without focusing on any specific technology or field.

I examine how variation in the words used in a patent reflect the technological direction of that patent. For example, by comparing the most frequent knowledge classes across patents that mitigate climate change, target cancer treatment or produce automation technologies, we can see how each purposes is reflected in the patent vocabulary. Figure B.5 shows the average weight for all knowledge classes split over three patent types. In Figure 2.7, I present three of the 50 estimated knowledge classes, and the average weight for patents of each type. Clearly patents which target cancer treatment use can be distinguished as using words such as *cell*, *antibody*, *gene*, while automation patents use *computer* and *information*. This figure supports the empirical concept of the knowledge space:the patent text is informative of the knowledge content of an innovation.

FIGURE 2.7
WORDCLOUDS AND KNOWLEDGE CLASS DISTRIBUTIONS BY PATENT TYPE



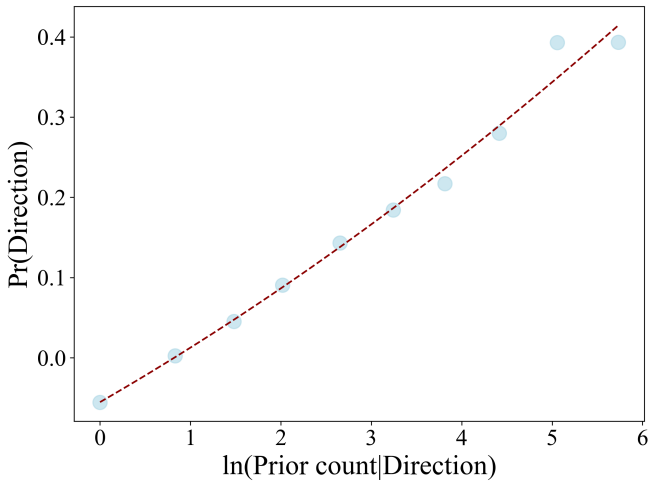
Notes: The bar chart shows the mean weight on a select three of the fifty knowledge classes, averaged across patents of each type. These types are not mutually exclusive. The word cloud is plotted using the estimated knowledge class to word distributions. The word size reflects the probability of using that word when describing that knowledge class.

I first show that prior knowledge shapes future innovations. Take the patent level count of prior work defined in equation 2.6 and include an additional condition to the indicator function: that $z_p = 1$. This will count the quantity of prior work in that field which achieves direction z. In Figure 2.8, I show that the probability a patent targets direction z is increasing in the quantity of prior work in that research field which also targets z. This is demonstrated more rigorously in Table B.1 where each additional patent increases the probability a patent targets direction z for between 2 to 4 percentage points.[27]

Importantly, Figure 2.8 shows no non-linear effect. This demonstrates two important features of endogenous growth. Prior work reduces the cost of future innovations, but also leads to path dependence. Path dependence refers to how the direction and nature of future innovation is determined by past work. Where early stage advances establish a path that is difficult to change. Aghion et al. (2016) show how changing the direction of a research field, for example to go green, is a challenge since the relative cost of producing either green or dirty

[27]In this table, and the later treatment model I stack the three directions into one regression model and include a period $\times$ direction fixed effect. This leads to an tripling of the sample size, and the effect is now the average across each direction. This achieves the goal of presenting technologically neutral results.

FIGURE 2.8
PATENT DIRECTION



Notes: This figure plots the probability that a patent achieves a given direction z by the log count of the number of prior patents existing in the local knowledge field of that patent, which also target z. The three directions are 1) mitigate climate change (PatentsView, 2024), 2) improve cancer treatment (Cancer Moonshot: USPTO, 2024) and 3) automate production (Mann and Püttmann, 2023). All three are stacked into one model.

patents is a function of what came before. This can be seen in Figure 2.8 as each successive patent targeting a given direction increases the chances of future work doing so further.

At the team level, this linear effect leads to straightforward outcomes. Here I use the same treatment as defined in section 2.4, however again adding the new condition that the patent belongs to the team's field, and targets direction z. Therefore the treatment now captures how many prior patents, targeting direction z, that the team loses following the premature death of a colleague. Again using the stacked regression model, I find that the probability a team's next patent targets a given direction z decreases by around 1 percentage point, for each prior-patent targeting the same direction removed. Naturally, following the premature death of an inventor, teams that lose access to the required knowledge to produce a patent of a certain type, see a change in the knowledge content of their innovations.

These results demonstrate that the death of a team member changes the innovation output of the team. Conditional on them returning to patent, they are therefore pivoted into new research fields.

TABLE 2.4
TEAM TREATMENT ESTIMATES: DIRECTION

	Dependent variable: Pr(Direction)			
	1.	2.	3.	4.
$D_{\tau t}$ | Direction	-0.0064***	-0.0150***	-0.0075***	-0.0083***
	(-9.45)	(-4.57)	(-3.71)	(-3.85)
Prior work$_{\tau_1 t}$ | Direction	0.0087***	0.0526***	0.0255***	0.0255***
	(21.90)	(12.78)	(13.56)	(13.58)
Volume $_{\tau}$				-0.5480*
				(-2.08)
N	91419	62487	62487	62487
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period × Direction FE		✓	✓	✓
Year × Direction FE			✓	✓

Notes: The first column uses a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each pair (τ_1, τ_2). The dependent variable is a stacked indicator for whether a patent achieves the given direction. The three directions are 1) mitigate climate change (PatentsView, 2024), 2) improve cancer treatment (Cancer Moonshot: USPTO, 2024) and 3) automate production (Mann and Püttmann, 2023). Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

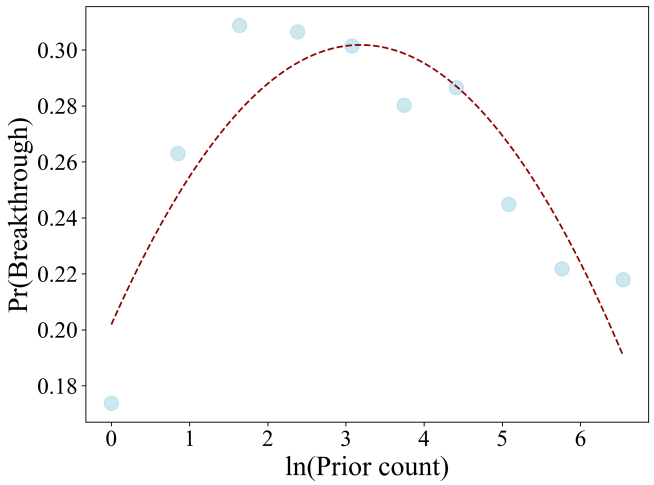
2.6.2 Breakthrough Innovations

I present the main results to test the hypotheses laid out in section 2.2 using the empirical strategy detailed in section 2.4.[28] I first show supporting evidence for Hypothesis 1. Figure 2.9 shows that the probability a patent becomes a breakthrough is an inverted-U shape in the quantity of prior work on which it builds. Recall the definition of a breakthrough patent using equation 2.7. Prior-count appears in the denominator, we would posit that the derivative be negative. However, we see that for low levels of prior work, this function is increasing. This is supporting evidence that prior work increases the impact of an innovation, and in fact post-count is determined in some part by what came before. However, the function later inflects as prior work becomes a barrier to

[28] I use the breakthrough measure defined endogenously by the knowledge space in equation 2.7. To remove concerns that this may be driven by some mechanical feature of the model I replicate all results using the Kelly et al. (2021) data in appendix section B.5.

novelty. As prior work accumulates, teams find it harder to be novel and this effect wins out, thus turning the slope back to the negative coefficient expected from the breakthrough definition.

FIGURE 2.9
PATENT BREAKTHROUGH



Notes: This figure plots a binned scatter plot and fitted regression line. The log count of the number of pre-existing patents in a patent's research field is split into 10 equally sized bins and the Y-axis plots the probability of a breakthrough within each bin. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent).

I test Hypothesis 2 at the team level. I first present the set of results averaging over all teams. All variables are defined as in section 2.4. All regression tables show a set of regression models that increase in rigour in each additional column. I interpret all results taken from the final column. Table 2.5 shows that on average, the novelty component of a breakthrough wins out, and teams benefit from moving to less explored research areas. This frees them from established paradigms, as they produce more breakthrough ideas.

To put these coefficients into tangible numbers, consider the following comparison. Each inventor contributes differently to the team. The justification for a continuous treatment model is that it matters who is lost, and which knowledge they contributed to the team. The average treatment measures the typical impact on a team's local field when a team member is lost. By estimating the average number of patents typically lost after such an event, I can predict how this change influences a team's ability to innovate.

TABLE 2.5
TREATMENT TEAM REGRESSION ESTIMATES

PANEL A: BREAKTHROUGH

Dependent variable: Pr(Breakthrough)				
	1.	2.	3.	4.
$D_{\tau t}$	0.0005**	0.0050***	0.0036**	0.0024*
	(3.25)	(3.35)	(3.16)	(2.00)
Prior work$_{\tau_1 t}$	-0.0008***	-0.0211***	-0.0122***	-0.0123***
	(-11.77)	(-6.14)	(-5.18)	(-5.03)
Volume$_{\tau}$				-2.3665***
				(-4.18)
N	30473	9498	9498	9498
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

Notes: The first column uses a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each pair (τ_1, τ_2). The dependent variable is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

For the sample of teams which return to patent without replacing the deceased inventor, the average number of patents lost from a team's local knowledge is 174. This change results in a 9.97 percentage point increase in the probability of producing a breakthrough, this represents a 22.21% increase on the baseline.[29] Although, for those that replaced the inventor, the average treatment was to only lose 43 patents. This leads to only a 5.47% increase on the baseline, given that on average teams close some of the gap.

[29]The change in probability is calculated using the baseline probability of 0.449. The coefficient of 0.0023 and an average treatment of 174 yield a change in log-odds of 0.4002. Applying this to the baseline and converting this back to probability gives 0.545, indicating a change of approximately 9.97 percentage points, which is a 22.21% increase relative to the baseline.

2.6.3 Heterogeneous Effects

Given the inverted-U shape in Figure 2.9, I show how this translates into a heterogeneous treatment effect. Figure 2.10 plots the same regression results, however split over four samples. I split the sample into quartiles of prior work in the initial team's knowledge field and run the model for each sub-sample. We see that the inverted-U shape translates directly into recommendations at the team level. For teams building on advanced areas, reducing the quantity of prior work by the average treatment of 174 patents increases their chances of a breakthrough by 41.79%. However, for those already working in early-stage research fields, the same change reduces their chances of a breakthrough by 60.83%. Importantly, the teams in early-stage areas which replace their inventor see a significantly smaller decrease in their ability to produce breakthroughs. If the number of patents lost is reduced to the average by replacing the inventor, these results show a much smaller 7.98% decrease in the likelihood of a breakthrough. This demonstrates the importance of the availability of knowledge within inventor networks.[30]

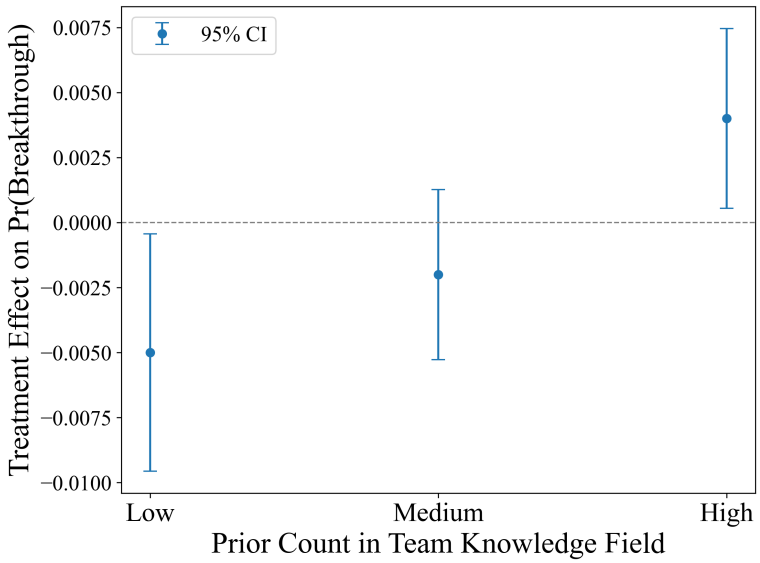
For teams in advanced areas, removing a team member who contributes the established knowledge increases their chances of producing a breakthrough. For them, increasing the novelty of their patents is key, and therefore moving into less-explored research fields improves their ability to be novel and produce breakthroughs. However, for a team in the first quartile, those that are already building on relatively little prior work, the same change is detrimental. If they remove a member who contributes the little knowledge on which they are building, their chances of producing a breakthrough reduce further. They move too far down the ladder of development and their innovations lose impact.

2.6.4 Novel Patents

To demonstrate the role of novelty, I again implement the data provided by Arts, Hou, and Gomez, 2021. They count the number of new words, bi-grams and tri-grams created by a patent. Where a bi-gram is the novel combination of two existing words, for example the first patent to introduce the term *artificial intelli-*

[30]This result suggests a valuable follow-on research project which studies frictions in the *market for collaborators*. If a team suffers the premature death of a collaborator who provided a certain type of required knowledge, perhaps there is a deficit in the supply of this knowledge, and they cannot be easily replaced. This variation in post-death team outcomes may be driven by their ability to find replacement inventors.

FIGURE 2.10
TREATMENT COEFFICIENT BY PRIOR-COUNT QUARTILE

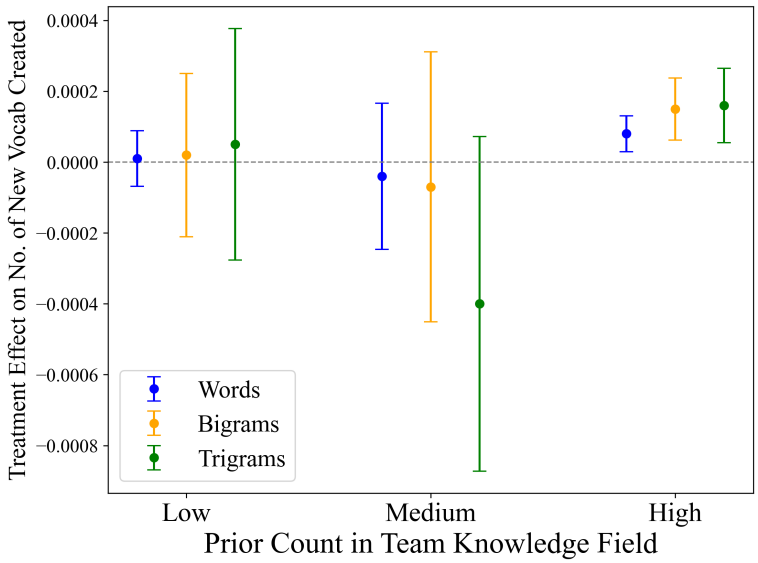


Notes: This figure plots the heterogeneous treatment effect for the continuous treatment variable outlined in equation 2.13. The x-axis plots the coefficient on the treatment for each of the four quartiles of the quantity of prior work in a team's knowledge field, prior to the premature death of a collaborator. The dependent variable is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Each model is a conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

gence to the USPTO vocabulary. I replace the outcome variable in the regression specification of equation 2.12 with the count of each new vocabulary type, and therefore run a fixed effect OLS regression model, instead of a logit. I again plot the coefficient on the treatment variable $D_{\tau t}$ for each of the outcome variables, by each of the four quartiles of $n_{\tau t}$.

As you can see in figure 2.11, for teams building on advanced areas (Q4), reducing the quantity of prior work they are building on leads them to be more novel. They build on less explored areas of the knowledge space, introduce more new vocabulary and create more breakthroughs. Interestingly however, teams building on low levels of prior-work (Q1-Q2), for which I estimate a negative treatment effect in figure 2.10, see no significant change in novelty. Given that breakthroughs require increasing both novelty and impact, holding novelty constant, this implies a decrease in impact. I argue that this is supporting evidence

FIGURE 2.11
TREATMENT COEFFICIENT BY PRIOR-COUNT QUARTILE: NEW VOCABULARY



Notes: This figure plots the heterogeneous treatment effect for the continuous treatment variable outlined in equation 2.13. The x-axis plots the coefficient on the treatment for each of the four quartiles of the quantity of prior work in a team's knowledge field, prior to the premature death of a collaborator. The dependent variable is the count of new vocabulary for new words, bi-grams and tri-grams taken from Arts, Hou, and Gomez, 2021. Each model is a fixed effect OLS model with team and patent order fixed effect models and standard errors are clustered at this level. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

for these teams in fact being on the upward sloping region of the inverted-U shape.

2.6.5 Robustness and Mechanisms

I examine the robustness of these results through two extensions. First, in Table B.9 I replace the breakthrough measure used in this paper with the comparable measure from Kelly et al. (2021). They also use text analysis to define patent similarity, where they define a breakthrough score comparing the similarity of a patent to the stock of knowledge that came before it and after it. I show that the same trend holds when using their measure, which reduces concerns that I may be capturing a mechanical effect from defining breakthroughs and the team span in one space.

On an alternative note, the innovation literature emphasises how combinatorial possibilities drive breakthrough innovations (Weitzman, 1998; Fleming, 2001;

Fleming and Sorenson, 2004; Singh and Fleming, 2010). Therefore one may worry that the positive coefficient given in Table 2.5 could be driven by the sample of teams that replace the inventor. If those teams bring in new knowledge and perspectives, thus increasing their combinatorial possibilities, this literature would expect their breakthrough chances to increase.

The team span represents the set of potential combinations of inventor knowledge profiles. If on replacing the deceased inventor the team expands the team span by increasing v_τ, this increases the recombination potential of the team.[31] Importantly, both hypotheses 1 and 2 are defined using the density of prior innovations in team's knowledge field.[32] Consult B.3 for further detail on these two hypotheses. Therefore this concern is in part dealt with by controlling for the volume of the team span in the empirical specification.[33] The coefficient on the treatment is estimated conditional on the mass of recombinations available to the team.

To examine this effect further I first run the empirical model on the sub-sample of teams which do not replace the deceased inventor, so that their team span decreases in size and they lose recombination possibilities.[34] These results are reported in Table B.6. On this sample the positive coefficient on the treatment D_τ remains, though it does lose power, and is now insignificant at the 5% level when controlling for team span volume. This is in part due to the much smaller sample size, as there are only 60 *no replace* teams compared to 447 *replace* teams.

I therefore use the larger sample of teams which do replace the deceased inventor to further test the mechanism at play. I examine whether the positive coefficient can be explained through teams bringing in new knowledge and expanding their set of recombinant possibilities, for which an increase in the vol-

[31] Formally, the team span contains an infinite set of points both before and after replacing a member. The increase is not in the cardinal number of possible combinations, but in the breadth of the space spanned by the team's knowledge.

[32] Defined as the count of prior innovations within their knowledge field, normalised by the volume of their span v_τ.

[33] I also run an additional robustness check in Table B.10 in which I show the reverse case to the treatment defined here. I run the same regression, on the same controls, but with a sample of teams that add an inventor, instead of removing one. I show that for both the breakthrough and direction model, the results flip their sign. To facilitate comparison of the coefficients I define $\mathrm{D}_{\tau t} = n_{\tau_2 - 1} - n_{\tau_1 - 1}$. Again here the variance is large, and the coefficients are not precisely estimated. However this is not due simply to a smaller sample, and warrants further study.

[34] To be precise, the volume decreases weakly since the deceased inventor knowledge profile have been an interior point of the team's knowledge field.

ume of their team span serves as a proxy. For this sample I run an additional interaction model, in which I introduce the interaction between the treatment D_τ and the team span volume v_τ. These results are reported in Table B.8.

First of all, for all specifications and treatment samples estimated, the coefficient on volume is negative and significant. For the within-team model, increasing the number of combinatorial possibilities does not increase the probability of breakthrough for a team. Second the coefficient on the treatment of reducing the number of previous innovations within a team's knowledge field remains positive and significant. In addition the interaction between the treatment and a team's volume is negative. This can be seen most clearly in Figure B.7. The treatment effect—capturing the team's move into a more novel, less explored area of the knowledge space—declines convexly over the volume of the team span. The shape approximates an inverse relationship. This provides strong supporting evidence for Hypotheses 1 and 2 and underscores the central mechanism: the density of prior work shapes the value of exploring new areas.

2.7 Conclusion

In this paper I ask how the development stage of a research field determines a team's ability to produce breakthrough innovations. A deeper understanding of the determinants of breakthroughs is key to modelling how the innovation frontier moves forward over time. Traditionally, the literature on knowledge production has focused on value. This paper presents a contribution to the innovation literature by constructing a unifying framework for teamwork capable of capturing the creation of new and successful research fields.

I model collaboration directly through the lens of a Bayesian model of Natural Language Processing, utilising the novel model of collaboration through text introduced in Chapter 1. I build a map of inventors, teams and patents in which to study how teams innovate. I refer to this as the knowledge space. As the first to integrate inventor teams and patents into one consistent space, the paper reconceptualises how knowledge is produced by recombining existing knowledge and standing on the shoulders of giants. The paper contributes a greater understanding of the key latent variables behind knowledge production and allows me to tackle a set of important hypotheses on which systematic evidence was missing.

The framework developed in this paper is required to back out a latent representation of a team's local knowledge field. The combination of the high-dimensionality of patent text data, and the computational Bayesian model allows me model teamwork in a tractable approach. I use premature inventor deaths to identify the effect of pivoting to more or less advanced research fields on a team's ability to produce breakthrough innovations. I find a non-linear relationship between prior work and breakthroughs. I find that teams produce more breakthroughs when building on enough prior work to incorporate valuable prior knowledge, but not so much that it stifles novelty.

The framework presented here marks the beginning of a rich future research agenda. The knowledge space provides a rich environment in which to study teams, but can be integrated with economic models to explain the broader innovation landscape. For example, modelling public R&D financing or firm innovation choices. Another key avenue for future work is to study the role of learning in this context and develop a dynamic version of the model. I hope that others are encouraged to utilise this framework to continue deepening our understanding of how we produce science and technology.

Chapter 3

From Shares to Machines: How Common Ownership Drives Automation

Abstract

Does increasing common ownership influence firms' automation strategies? We develop and empirically test a theory indicating that institutional investors' common ownership drives firms employing workers in the same local labor markets to boost automation-related innovation. First, we present a model integrating task-based production and common ownership, demonstrating that greater ownership overlap drives firms to internalize the impact of their automation decisions on the wage bills of their local market competitors, thereby fostering more automation and reducing employment. Second, we empirically validate the model's predictions. By analyzing patent texts, the geographic distribution of firms' labor forces at the establishment level, and exogenous increases in common ownership due to institutional investor mergers, we isolate the effects of rising common ownership within and across labor markets. Our findings reveal that firms experiencing a positive shock to common ownership with labor market rivals exhibit increased automation, coupled with a decrease in employment. Conversely, similar ownership shocks do not lead to heightened automation innovation if firms do not share local labor markets.

This chapter was co-authored with Dennis C. Hutschenreiter, Felix Noth, Stefano Manfredonia and Tommaso Santini. I gratefully acknowledge the support of the Spanish Agencia Estatal de Investigación (MCIN/ AEI /10.13039/501100011033) through grant PID2020-114251GB-I00.

3.1 Introduction

Does institutional investors' common ownership affect the direction of technological progress, innovation, and automation strategies of their portfolio companies? In this paper, we show that common ownership, i.e., the overlap of the shareholder base of public corporations, leads portfolio firms operating in the same local labor markets to increase their innovation with the intent of automating their production, with important implications for employment outcomes.

Common ownership of publicly traded firms and the automation of tasks previously performed by workers are both rising phenomena in developed economies. Backus, Conlon, and Sinkinson, 2021 build a measure of common ownership and document that it has tripled among the firms in the S&P 500 between 1980 and 2017. Over the same period, the 10 largest institutional investors have quadrupled their ownership of U.S. stocks and, by the end of 2016, they managed 26.5% of total equity assets (Ben-David et al., 2016). Economic theory suggests that common ownership of firms competing in the same product market can reduce competition, pushing such markets toward monopolistic outcomes, with consequences for consumer welfare.[1]

On the other hand, growing concerns have emerged regarding the impact of automation technologies on employment, welfare, and inequality. These concerns have been fueled by recent technological advancements, predictions of future developments, and the increasing adoption of automation technologies across various sectors (Frey and Osborne, 2017; Arntz, Gregory, and Zierahn, 2016). Numerous studies have explained the stagnation of median real wages and the decline in wages for less-educated workers from a macroeconomic perspective, attributing these trends to the rise of automation (Acemoglu and Restrepo, 2018; Moll, Rachel, and Restrepo, 2022; Santini, 2024). Additionally, studies focusing on local labor markets have identified negative effects of automation—proxied by robot adoption—on employment and wages (Acemoglu and Restrepo, 2020; Dauth et al., 2019). However, when using firm-level data, the evidence about the effect of robot adoption on employment and wages is mixed. Some studies find a positive association (Koch, Manuylov, and Smolka, 2021; Deng et al.,

[1] See, for example, Macho-Stadler and Verdier, 1991, Baker, 2015, Posner, Scott Morgan, and Weyl, 2016, Backus, Conlon, and Sinkinson, 2021, Anton et al., 2018. Similarly, Azar, Raina, and Schmalz, 2022 and Azar, Schmalz, and Tecu, 2018 present evidence that common ownership might lead to anti-competitive behavior, higher prices, and lower output in the airline and banking industries.

2024; Aghion, Van Reenen, and Zingales, 2013), while others find a negative effect (Bonfiglioli et al., 2024; Bessen et al., 2023).

Our paper aims to contribute to these two alternative strands of the literature and provide a better understanding of the incentives of firms to automate production. More specifically, we demonstrate the consequences of increasing common ownership of public corporations on automation innovation and employment outcomes through a labor market perspective.

In a task-based model of automation, we show that a firm experiencing an increase in their common ownership with rivals in local labor markets will increase the share of automated tasks. This is the case since firms with labor market power internalize the effect of their automation efforts on the wage bill of their commonly-owned rivals incentivizing them to reduce the labor demand.

We empirically test the model's prediction about the effect of common ownership within local labor markets on automation. To address potential endogeneity coming from automation-oriented investment strategies of institutional investors, we use mergers between institutional investors as quasi-natural experiments to exploit exogenous changes in common ownership. As it has been argued in previous literature (Lewellen and Lowry, 2021; He and Huang, 2017a), mergers increase common ownership and are unlikely to be motivated by policies or the performance of individual portfolio firms.

Since firms experience increases in common ownership due to mergers of institutional investors often several times throughout our sample, we apply the state-of-the-art *difference-in-differences* (DID) methods developed by De Chaisemartin and d'Haultfoeuille, 2024. Additionally, we developed a continuous treatment framework for common ownership since mergers of institutional investors affect firms heterogeneously. Moreover, in this setup, establishment-level information on the distribution of a firms labor force allows us to disentangle the causal effects of common ownership on our outcome variables, separately for scenarios with and without labor-market rivalry between portfolio firms. Therefore, this provides compelling evidence that our proposed mechanism is indeed in effect: common ownership increases automation if and only if firms interact in local labor markets.

To measure automation, we apply the classification of patents into automation and non-automation patents proposed by Mann and Püttmann, 2021. They identify automation patents from the textual content of each patent document. They

train a naïve Bayes classifier on patent texts, and classify the universe of USPTO utility patents from 1976-2014 as automation patents, or not. This is yet another example of the power of patent texts in describing the direction of innovation. As discussed by the authors, this approach outperforms those suggested in earlier studies that depend on indirect indicators like the proportion of routine tasks in job descriptions (Autor, Levy, and Murnane, 2003; Goos and Manning, 2007; Autor and Dorn, 2013), or on limited measures of automation such as expenditure on computer capital (Beaudry, Doms, and Lewis, 2010; Michaels, Natraj, and Van Reenen, 2014; Akerman, Gaarder, and Mogstad, 2015), or investment in robotics (Graetz and Michaels, 2018; Acemoglu and Restrepo, 2020). Furthermore, applying the Mann and Püttmann, 2021 classification, Danzer, Feuerbaum, and Gaessler, 2024 show that positive shocks to labor supply due to immigration lead firms to reduce automation innovation. At the same time, the effect on non-automation innovation is nil.[2] This underlines that patents classified as automation innovation capture firms' incentives to invest in labor-saving technologies.

We find that firms that experience an increase in common ownership with other firms operating in the same local labor market (i.e., in at least one shared commuting zone) increase patent output related to automation technologies. Simultaneously, we document a decrease in employment for these firms. In contrast, the effect of common ownership of firms operating in distinct labor markets on firms' automation innovation is not statistically significant. Hence, our empirical results suggest, that common ownership by institutional investors among labor-market competitors steers the direction of technological progress into more automation-related innovation, consistent with our theoretical model. Our paper sheds light on the relationship between corporate governance, labor-market competition, and automation.

A battery of tests is conducted to ensure the robustness of our results. First, we use alternative measures of the automation content of firms' innovation output by weighting patents by their truncation-adjusted citation counts (Hall, Jaffe, and Trajtenberg, 2001; Atanassov, 2013). Using these innovation measures, we corroborate our result that common ownership between labor market rivals increases innovation output related to automation, while non-automation innovation output is not affected. Second, we show that our results are robust to

[2] See also Terry et al., 2024 who find a positive impact of immigration on innovation in general.

sample selection pooling our two treatment setups: mergers of institutional investors that increase common ownership within and across local labor markets. Third, Lewellen and Lowry, 2021 suggests that the Global Financial Crises could drive the effects attributed to common ownership, as at the same time many firms have been affected by mergers of their institutional owners. Therefore, we corroborate our results using only data up to 2006. Finally, we use a traditional binary treatment variable in our difference-in-differences setting. We find the same qualitative result in all these tests: common ownership between labor market rivals boosts firms' automation-related innovation output.

Related Literature Our paper contributes to different strands of the literature in economics and finance. First, it contributes to the debate on the impact of common ownership on the firm's objective function and resulting behavior. The effect of increasing common ownership on product market competition and consumer welfare, as well as its implications for antitrust policy, has been investigated by academics in recent years (Baker, 2015; Posner, Scott Morgan, and Weyl, 2016; Azar, Raina, and Schmalz, Backus, Conlon, and Sinkinson, 2021). Concerning innovation, López and Vives (2019) show that common ownership may increase R&D investments if it leads firms to internalize the positive externalities of technology spillovers on product market rivals, and Anton et al. (2018) present evidence that common ownership on the firm-pair level might have either positive or negative effects on innovation depending on the relative degrees of technology spillovers and product market rivalry between the firms (Bloom, Schankerman, and Van Reenen, 2013). Finally, Hutschenreiter (2023) shows that common ownership leads to higher technology diffusion across portfolio firms. We contribute to this literature by investigating how common ownership affects another dimension of firms' innovation strategy, namely the automation content of their innovation output. We further document a labor-market channel and a firm-level reduction in employment growth due to common ownership.

Another pertinent line of research closely related to our paper lies in the intersection of common ownership, labor market dynamics, and automation. Azar and Vives (2019, 2021) study the effects of common ownership on income shares of production factors in a general equilibrium model, but do not consider automation. Azar, Qiu, and Sojourner (2022) study the effect of common ownership concentration on local labor-market outcomes and Azar et al. (2023) examine the relationship between monopsony power and automation adoption. We con-

tribute to this literature by presenting firm-level evidence on the relationship between common ownership and the automation-related outcome of innovation strategies. Furthermore, our unique estimation strategy allows us to identify the mechanism behind this relationship: labor-market rivalry is a necessary condition for common ownership to spur investment in automation innovation. That is, we can disentangle the effect of common ownership on our outcome variables for firms operating within and across local labor markets. Furthermore, using the setup of institutional mergers as proposed by Lewellen and Lowry (2021) allows us to present causal estimates applying state-of-the-art dynamic difference-in-difference methodology (De Chaisemartin and d'Haultfoeuille, 2024).

Finally, our paper is related to research on the impact of automation on wages and employment. Initiated by the seminal research of Acemoglu and Restrepo (2020) for the U.S. and followed by Dauth et al. (2019) for Germany,[3] both studies find negative effects of robot adoption on employment and wages using a local labor market approach. Afterward, the literature transitioned to utilizing firm-level data. This later development presents the challenge of establishing causality by identifying credible exogenous variations. Studies by Bonfiglioli et al. (2024), Bessen et al. (2023), and Aghion et al. (2020) have addressed this issue. The first two papers report negative employment effects, while the third finds a positive effect, arguably due to the different automation proxy used—specifically, investment in industrial equipment—which is likely more complementary to labor.

Finally, several studies examine the firm-level outcomes following the adoption of robots. Deng et al. (2024) for Germany, Koch, Manuylov, and Smolka (2021) for Spain, and Acemoglu, Lelarge, and Restrepo (2020) for France all find that employment *increases* in firms after the adoption of industrial robots. We contribute to the literature by identifying an additional mechanism that leads firms to increase their automation effort. That is, we show that a part of the surge in firms' investments in automation technologies is the result of common ownership among local labor-market rivals. Common ownership leads firms to internalize the negative externality of employing workers on the rivals' wage bill. Thus, common ownership in local labor markets increases the incentives to invest in innovation that allows the firms to save labor through the automation

[3] In the German context, Dauth et al. (2019) find that robot adoption decreases employment in the manufacturing sector while increasing it in the service sector, keeping aggregate employment unaffected. This mechanism has been formalized by Hutschenreiter, Santini, and Vella (2022).

of tasks. Hence, automation that is driven by common ownership instead of other reasons such as firm growth or improved productivity could lead to a more negative relationship between automation and labor-market outcomes. These include wages, employment, and the labor share of income which could exacerbate the problem of "*excessive automation*" (Acemoglu, Manera, and Restrepo, 2020).

Paper Outline The rest of the paper is organized as follows. Section 2 outlines the theory and proves a set of theoretical propositions. Section 3 presents the data, identification strategy and empirical results. Section 4 concludes.

3.2 Theory

In this section, we present a simple model that we employ to derive testable empirical hypotheses. Mathematical derivations are relegated to the appendix section C.1.

3.2.1 Theoretical Model

Consider an economy with J firms. We call one of these, firm f, the focal firm, and analyze its automation strategy. The firms operate their production processes in a set of local labor markets $\mathbf{C}$, which we interpret as the collection of commuting zones.[4] Thus, a firm j executes its production in a set $\mathbf{C_j} \subset \mathbf{C}$. We say that a firm j has local labor-market (LLM) overlap with the focal firm f if both employ production plants in some local labor market at the same time, i.e., there exists some location $c \in \mathbf{C_f} \cap \mathbf{C_j}$.

Given the geographic distribution of firm f's production sites, we can partition the set of the remaining $J-1$ firms in the economy into two subsets. Namely, the set $\mathbf{R_f}$, such that a firm $j \in \mathbf{R_f}$ has LLM overlap with firm f, and the set $\mathbf{N_f}$, such that $j' \in \mathbf{N_f}$ implies that $\mathbf{C_f} \cap \mathbf{C_{j'}} = \emptyset$. Moreover, we define $\mathbf{R_f^c} \equiv \{j | j \in \mathbf{R_f}, c \in \mathbf{C_j}\}$, the set of all firms $j \neq f$ that operate a plant in a location $c \in \mathbf{C_f}$ in which firm f is also present.

We assume that firm f has some degree of *labor market power* in the local labor markets $c \in \mathbf{C_f}$ in which it is present. To model this most simply, we assume that the focal firm f has full knowledge of the labor supply structure and takes into

[4] To match our empirical analysis, we consider the geographic distribution of firms' production plants as exogenous.

account that $\partial w_j^c / \partial L_f^c = \rho_{jf}^c > 0$ for all $j \in \mathbf{R_f} \cup \{f\}$, $c \in \mathbf{C_f} \cap \mathbf{C_j}$, where L_f^c is firm f's labor demand in c and w_j^c the wage firm j has to pay in order to employ a given amount of labor in the same location. For instance, if firm f increases its labor demand in some of its plants it increases the outside option of workers in the locations in which these plants operate. Thus, we call the firms in $\mathbf{R_f}$ firm f's labor-market rivals.

We abstract from wage spillover effects between LLMs, i.e., $\partial w_{j'}^{c'} / \partial L_j^c = 0$, for all $j, j' = 1, 2, \ldots J$ and all $c, c' \in \mathbf{C}$, $c \neq c'$. In particular, this implies that firm f cannot affect the wages that other firms pay in locations $c' \notin \mathbf{C_f}$, in which it does not operate, i.e., $\partial w_j^{c'} / \partial L_f^c = 0$, for all $c \in \mathbf{C_f}$, $j \in \mathbf{N_f}$. Hence, there is no labor-market rivalry between firms in $\mathbf{N_f}$ and f.

For simplicity, we abstract from product market competition and all firms are price takers in the capital market.[5] We can think about the firms $j = 1, 2, \ldots, J$ as multi-product firms producing different goods Y_j^c in each establishment and selling them to a global market at a given price p_j^c, such that they do not have price-setting power in their respective product markets. They take the rental rate r of capital K as given.

There exists a collection of institutional investors who may own shares in the firms. Drawing on the literature, which suggests that good corporate governance induces management to maximize a weighted average of investors' cash flows from their portfolio, we posit that firm f's objective function, under common ownership, internalizes the impacts of its strategic decisions on the profits of other portfolio firms. As shown by López and Vives, 2019, we can thus write firm f's objective function as

$$\phi_f = \pi_f + \sum_{j \neq f} \lambda_{fj} \pi_j \tag{3.1}$$

where π_j is the profit function of firm j and $\lambda_{fj} \geq 0$ is the profit weight firm f puts on firm j's profits. The parameter λ_{fj} is a function of the cashflow rights of firm f's investors to the profits of firms f and j. In particular, λ_{fj} increases in the degree of ownership overlap of the two firms. Thus, an increase in common ownership between the two firms is modeled as an increase in λ_{fj} in our analysis.

[5] For a model with product market competition, see Hutschenreiter and Santini, 2021, in which the effect of common ownership on automation depends on the ratio of factor supply elasticity. In the case in which capital supply is more elastic than labor supply, common ownership leads to an increase in automation.

At each location $c \in \mathbf{C_f}$, focal firm f has access to a technology that by performing a continuum of distinct tasks $x^c \in \mathbf{X}^c = [0,1]$ produces output Y_f^c. The final output of firm f in location c is given by the production function

$$Y_f^c = \exp\left(\int_{\mathbf{X}^c} \ln\left[y_f(x^c)\right] dx^c\right)^{\nu} \tag{3.2}$$

where $y_f(x)$ is the quantity of the task (indexed by x^c) employed in production. Each amount of task $x^c \in \mathbf{X}^c$ performed in a location $c \in \mathbf{C_f}$ is produced according to the following intermediary production function,

$$y_f(x^c) = \gamma_m^c(x^c) m_f(x^c) + \gamma_\ell^c(x^c) \ell_f(x^c), \tag{3.3}$$

in which quantities of machines $m_f(x^c)$ and labor $\ell_f(x^c)$ are perfect substitutes, and $\gamma_m^c(x)$ and $\gamma_\ell^c(x)$ are the productivity schedules of capital and labor over the task measure. Without loss of generality, we assume that at each production site of firm f in the locations, $c \in \mathbf{C_f}$, the productivity schedules are continuously differentiable over each set $\mathbf{X}^c$ and this set is ordered in such a way that the comparative advantage of producing a task with labor strictly increases in x^c, i.e., $d/dx^c(\gamma_\ell^c(x^c)/\gamma_m^c(x^c)) > 0$ for all $x^c \in \mathbf{X}^c$.

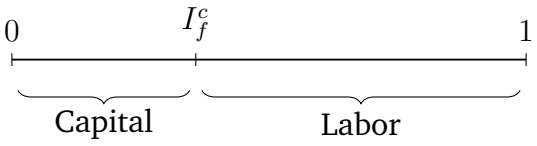
Then, firm f chooses its inputs, such that each set $\mathbf{X}^c$ of tasks is divided into two regions: the tasks produced with capital and the tasks performed by labor, as shown in Figure 3.1.

The threshold that separates the two sets is $I_f^c \in \mathbf{X}^c$. Then, $I_f^c \in [0,1]$ is the degree of automation of the plant in location c. Thus, firm f's average degree of automation is given by

$$I_f \equiv \frac{1}{|\mathbf{C_f}|} \sum_{\mathbf{C_f}} I_f^c \tag{3.4}$$

FIGURE 3.1

CAPITAL AND LABOR ALLOCATION OVER TASKS



Allocation of capital and labor to tasks over $\mathbf{X}^c$ in firm f's plant in location c.

To maximize its objective function firm f solves the following program.

$$\mathcal{P}_f \begin{cases} \max_{\mathbf{\Omega_f}} & \sum_{\mathbf{C_f}}(p_f^c Y_f^c - rK_f^c - w_f^c L_f^c) + \sum_{j\neq f}\lambda_{fj}\Big\{\sum_{\mathbf{C_j}}(p_j^c Y_j^c - rK_j^c - w_j^c L_j^c)\Big\}, \\ & \text{subject to} \\ & Y_f^c = \exp\left(\int_{\mathbf{X^c}} \ln\left[y_f(x^c)\right] dx^c\right)^{\nu} \\ & y_f(x^c) = \gamma_m^c(x^c) m_f(x^c) + \gamma_\ell^c(x^c)\ell_f(x^c) \\ & K_f^c = \int_{\mathbf{X^c}} m_f(x^c) dx^c \\ & L_f^c = \int_{\mathbf{X^c}} \ell_f(x^c) dx^c, \end{cases}$$

where $\mathbf{\Omega_f} = \{Y_f^c, K_f^c, L_f^c, y_f(x^c), m_f(x^c), \ell_f(x^c)\}$ and we define $L_f \equiv \sum_{\mathbf{C_f}} L_f^c$ as its total labor input. Moreover, $K_f \equiv \sum_{\mathbf{C_f}} K_f^c$ is the amount of capital it employs.

We are interested in the relationship between the degree of common ownership, λ_{fj}, with some other firm j in the economy and the optimal level of automation I_f. We state our first result in the following proposition:

Proposition 1. *If common ownership, i.e., the profit weight λ_{fj}, of firm f with respect to some labor-market rival firm $j \in \mathbf{R_f}$ increases, then the optimal level of automation I_f of firm f increases.*

Proof. See the theoretical appendix C.1. □

Intuitively, as common ownership with a labor-market rival increases, the extent to which firm f internalizes the profit of this firm also increases. The only way in which firm f can affect the profit of firm $j \in \mathbf{R_f}$ in our model is by decreasing its wage bill and, to achieve this it decreases the level of labor input in locations $c \in \mathbf{C_f} \cap \mathbf{C_j}$. That is firm f trades off the cost of reducing its labor demand with the benefit of decreasing firm j's wage bill, which it now internalizes to a higher degree. This implies a reduction in labor input by firm f in these plants. To mitigate the cost of this reduction, it chooses a higher degree of automation I_f^c in plants $c \in \mathbf{C_f} \cap \mathbf{C_j}$. In other words, an increase in common ownership increases the *internalized* marginal factor cost of labor for firm f, i.e., $(\partial w_f^c/\partial L_f^c)L_f^c + w_f^c + \lambda_{fj}(\partial w_j^c/\partial L_f^c)L_j^c$ in locations $c \in \mathbf{C_f} \cap \mathbf{C_j}$, that accounts for the cost of a marginal increase in labor demand by firm f on firm j's profits. Therefore, in each affected location the incentives to substitute labor with capital increase, increasing the set of tasks $[0, I_f^c]$ produced with capital by shifting I_f^c

to a higher indexed task $x^c \in \mathbf{X}^c$. The degree of automation I_f^c of the plants in $c \in \mathbf{C_f} \cap \mathbf{C_j}$ increases, leading to a higher degree of automation I_f of firm f. Hence, if the firm experiences a positive shock to common ownership with one of its labor-market rivals and it is producing at the ex-ante optimal level of automation, we expect that this shock causes the firm to increase its effort to automate additional tasks.

We have seen that common ownership in our model causes an increase in the optimal degree of automation if and only if a firm can affect the wage bill of the other firm through its labor demand. The next result follows immediately.

Corollary 1. *Common Ownership between firm f and a firm $j' \in \mathbf{N_f}$ does not affect firm f's optimal degree of automation I_f.*

Proof. The result immediately follows from the fact that firm f cannot influence the wage bill of firm j'. Thus, the profits of firm j' do not depend on firm f's strategic choices. □

In the next subsection, we discuss the results derived from the model to develop testable hypotheses.

3.2.2 Hypothesis Development

In this section, we translate the theoretical results into testable empirical hypotheses. Drawing on Proposition 1, we have seen that a necessary condition for common ownership to alter firms' strategies regarding automation is that firms have some degree of market power in the labor market. In particular, our mechanism requires that the focal firm, whose automation choice we observe, can influence the wage bill of the other portfolio firms with which it shares common owners. Therefore, in our empirical analysis, we will distinguish between firms that interact in local labor markets and those that operate in distinct labor markets. To this end, we define labor market rivals using their concurrent employment in the same commuting zones. In particular, we say that firms are local labor market competitors if they both have positive employment in establishments that are located in at least one shared commuting zone at the same time. Furthermore, we use a classification of patents into automation and non-automation patents based on patent texts to measure the automation content

of firms' innovation output. Then, we focus on a positive shock to the common ownership of a focal firm with respect to labor market rivals. Given the result in proposition 1, we test the following hypothesis:

Hypothesis 3. *An increase in common ownership with local labor market rivals causes the focal firm to increase the automation content of its innovation output.*

From our model, we expect to observe an increase in automation if the overlap in ownership between firms in the same local labor markets increases. Thus, after such a shock to common ownership occurs, a firm has to adapt its degree of automation by innovating. Furthermore, as Corollary 1 states, we expect the effect of common ownership to be absent, if we consider the overlap in ownership of the focal firm with those firms in the investors' portfolios unaffected by the focal firm's labor demand decisions. Thus, we expect that an increase in common ownership among firms that do not operate in the same labor markets does not increase the automation content of firms' innovation output. Hence, we test the following hypothesis:

Hypothesis 4. *An increase in common ownership with firms employing labor in different commuting zones does not cause the focal firm to focus more on automation innovation.*

3.3 Empirical Analysis

We bring both of these hypotheses to the data to test the implications derived from the model. In this section, we discuss the data sources, the variables that we utilize, and our identification strategy. Finally, we report the empirical findings.

3.3.1 Data Sources

We build a novel data set that combines seven different data sources: (i) We start by retrieving firms' financial information from COMPUSTAT; (ii) we merge this information with the number of outstanding shares and stock prices from CRSP; (iii) Thomson Reuters form 13F file provides firms' institutional shareholder information, i.e., the institutional investors and the number of outstanding shares owned by each of them; (iv) We gather data on the geographic distribution of

firms' labor force from the establishment-level NETS database for each county and map them to US commuting zones that we define as local labor markets; (v) We use patent information from the USPTO and the DISCERN database (Arora, Belenzon, and Sheer, 2021) that provides us with a match of patents to public corporations; and (vi) We obtain M&A data from Lewellen and Lowry, 2021 for mergers between institutional investors. Finally, (vii) we use the Mann and Püttmann's (2021) classification of the universe of USPTO patents as either automation or non-automation patents. We now explicitly define the variables used in our empirical analysis.

3.3.2 Variables

Common Ownership

To measure common ownership, we follow the recent literature and use the *Cindex* (Lewellen and Lowry, 2021). This measure of common ownership is symmetric (undirected). Common ownership of firm j at time t is defined in the following way:

$$\textit{Cindex}_{jt} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i} \beta_{ij} \beta_{ik}, \tag{3.5}$$

where j denotes the focal firm, i indicates investors owning it, and $k = 1, 2, \ldots, K$ indexes the relevant set of firms. Furthermore, β_{ij}, β_{ik} are the fractions of outstanding shares of firms j and k, respectively, that investor i owns as of the calendar year-end. Because we want to study the effect of changes in common ownership of a focal firm with its local labor market (LLM) rivals, we construct our main measure of common ownership $\textit{CindexLLM}_{jt}$, such that it only takes into account these K_{LLM_j} firms that have labor-market overlap with firm j, that is they operate in at least one commuting zone in which the focal firm j is present. Moreover, we also compute $\textit{CindexALL}_{jt}$, for which the relevant set of firms is K_{ALL}, i.e., all firms.

Automation Patents

For our empirical analysis, we need a dynamic measure of automation at the firm level. To this end we use the data provided by Mann and Püttmann, 2021. This paper defines automation as a *"device that carries out a process independently"*. Using this definition they train a classification model on patent texts to classify all USPTO patents, awarded between 1976 and 2014, as either automation or

not.[6] Table 3.1 gives examples taken from the original paper of both innovation types.

In summary, the authors train a naïve Bayes classifier. They define two classes, automation and non-automation. They first manually classify 560 patents into both classes and define, through the mutual information criterion, a set of words informative about each class. Words that identify that a patent is an automation device are *automat*, *output*, *execut*, *inform*, *input* and *detect*.[7] The algorithm then uses the occurrence of these words to calculate a probability that the patent belongs to each class and, by taking the maximum, they classify each patent.

The authors present the standard validation tests. In the training sample, the algorithm and manual coding agree 80% of the time, and the probability of a false positive (type-1) and a false negative (type-2) are 21 and 17 percent, respectively. For the out-of-sample testing, performance declines slightly. However, the corresponding statistics for true positives, false positives, and false negatives are 77%, 23%, and 22%.

TABLE 3.1

AUTOMATION PATENT EXAMPLES

Patent title	Patent number	Automation?
"Automatic taco machine"	5531156	Yes
"Automated email activity management"	7412483	Yes
"Hair dye applicator"	6357449	Yes
"Bicycle frame with device cavity"	7878521	No
"Process for making pyridine-thione salts"	4323683	No
"Golf ball"	4173345	No

This table shows examples of patents' classification into automation or non-automation patents. Source: Mann and Püttmann, 2021.

Using a broad definition of innovation and the machine learning method allows us to measure innovation for a large sample of firms, across industries and time. While errors in the classification process introduce noise, as long as there is no

[6] The paper restricts the sample to utility patents and therefore does not classify design patents, however, as these do not "carry out" a process their comparison would introduce noise in our analysis.

[7] These words are stemmed first, to capture variations of the same word, for example, automation, automate, and automatic all stem from the word automat.

systematic bias in the occurrence of type-1 and type-2 errors, we believe the measure provides a valuable source of information on firm automation strategies.

Given this patent-level classification, we construct different measures of innovation output on the firm-year level that allow us to evaluate changes in the automation strategy of firms. The first and main measure gauges the automation content of firms' innovation output, i.e.,

$$\textit{AutoRatio} = \ln\left[\frac{1 + \text{Number of automation Patents}}{1 + \text{Number of non-automation Patents}}\right] \tag{3.6}$$

AutoRatio measures the extent to which a firm focuses on innovation for automation vs. inventions unrelated to automation. Thus, it allows us to observe changes in the automation strategies of firms on the intensive margin.

The second measure, *AutoDummy*, is an indicator variable that takes the value one if a firm in a given year applies for a patent that was eventually granted and is classified as an automation patent according to Mann and Püttmann, 2021, and zero otherwise. We use this measure to assess changes in the propensity of firms to invest in automation technologies.

We also use patent counts to assess the effect of common ownership on automation and non-automation innovation separately to study whether changes in *AutoRatio* result from increases (decreases) in the number of automation (non-automation) patents, respectively. *lnAuto* and *lnNonAuto* are the natural logarithms of (one plus) the number of automation and non-automation patents, respectively.

Employment Growth

Since our model predicts that common ownership increases automation through a labor market channel, we also test whether common ownership leads firms to change their hiring behavior. We measure the growth rate of firm-level employment, that is,

$$EmpGrowth_t = \frac{Employees_t - Employees_{t-1}}{Employees_{t-1}}, \tag{3.7}$$

where $Employees_t$ is the number of employees (in thousands) of a firm in year t. Moreover, we compute the indicator variable *EmpIncrease* that takes the value

one if a firm experiences positive employment growth, and is zero otherwise.

Treatment Variables

Following the recent literature on common ownership and the estimation of its causal effects (Lewellen and Lowry, 2021), we use exogenous changes in common ownership due to the mergers of institutional investors. We use the information on 53 institutional mergers from 1990 to 2010 and, in particular, their announcement dates and the merging parties' ownership in the universe of publicly traded companies in the quarter before the announcement date to define a set of continuous and discrete treatment variables.[8]

Treated by merger—First, we define a set of firm-years in our panel that is treated by a merger of institutional investors similar to He and Huang, 2017b and Lewellen and Lowry, 2021. We call this set $\mathbf{T}$. The firms in $\mathbf{T}$ are treated in the sense that the merger is likely to increase their shareholder overlap with other firms. For this reason, we require that (i) for a firm (say firm 1, or the focal firm) to be treated by a merger, one of the merging investors (say investor A) holds at least 1% of outstanding shares of this firm before the merger (i.e., as of the quarter preceding the quarter of the merger announcement date). (ii) There must be at least one other firm (say firm 2) such that the other merging investor (say, B) owns 1% (again as of the quarter before the merger announcement) of this second firm. Furthermore, (iii) for this pair (firm 1, firm 2), neither of the two investors (A or B) can hold more than 1% in both firms in the quarter before the merger announcement, such that the merger is likely to lead to a new common shareholder (the merged institution) that holds at least 1% in both companies but did not do so before the merger. The firms satisfying these three criteria are likely to experience an increase in common ownership with some other firm in the economy due to the merger.[9]

[8] We checked the robustness of our results in a sample of firms from 1990 to 2006, excluding the last seven mergers in the sample provided by Lewellen and Lowry, 2021, because of concerns that these mergers and firm outcomes may be contaminated by the financial crisis. However, all our results stay qualitatively the same.

[9] At the same time, these firms are not likely to experience changes in their shareholder composition and concentration, since only one of the merging parties (A) holds more than 1% of outstanding shares, while we require the other investor (B) to hold less than 1% or none of the shares. This is crucial since Guo et al., 2024 have shown that mergers of institutional investors that both hold significant shares in one firm increase block holder ownership, which results in changes to firms' innovation strategy and outcomes. Because of the careful construction of our treatment sample and since we also distinguish the effect of treatment by mergers for firms within and across commuting zones, finding differential results, we can confidently conclude that changes in ownership concentration do not drive our results.

Treatment within Commuting Zones—Second, we define a subset $\mathbf{T}_{LLM} \subset \mathbf{T}$ of firm-years identified as treated by a merger above to construct our main treatment variable. We are interested in exploiting exogenous changes to common ownership between firms that operate in the same local labor markets. To do so, we modify criterion (ii) above such that we require the existence of a firm 2 that operates in at least one commuting zone in which also focal firm 1 operates, i.e., they have local labor market (LLM) overlap. Thus, firms treated in this sense likely experience positive changes to their common ownership with a local labor market peer. Hence, we expect that their *CindexLLM*, as defined above, increases.

For those firms that satisfy these three criteria, with LLM overlap, our first discrete treatment variable, *TreatLLM* takes the value one in the year of the quarter that immediately precedes the merger announcement and is zero otherwise.

Since firms may be treated to a different extent depending on the size of the holdings of the merging parties, we also construct a continuous treatment variable. This variable corresponds to the implied change in the *Cindex* of the focal firm to the other. To this end, we can compute the firm-pair level $\textit{Cindex}_{12} = \sum_i \beta_{i1}\beta_{i2}$ for a firm-pair (the focal firm 1 and its LLM rival, firm 2) at the time of the quarter preceding the merger announcement. We also can compute the counterfactual

$$\textit{Cindex}_{12}^{merged_{AB}} = (\beta_{A1} + \beta_{B1})(\beta_{A2} + \beta_{B2}) + \sum_{i \notin \{A,B\}} \beta_{i1}\beta_{i2}, \tag{3.8}$$

in which we treat the two investors as having already merged, using the same pre-announcement quarter ownership shares. The difference between the counterfactual and the actual *Cindex* is then given by

$$\Delta_{12}^{merged_{AB}} \equiv \textit{Cindex}_{12}^{merged_{AB}} - \textit{Cindex}_{12} = \beta_{A1}\beta_{B2} + \beta_{B1}\beta_{A2}, \tag{3.9}$$

that is the expected change of the firm-pair level *Cindex* due to the merger affecting the investors A and B in firms 1 and 2, and the β_{ik} are the corresponding holdings of the investors $i \in \{A, B\}$ in firm $k = 1, 2$ as of the quarter before the merger announcement. The firm-level continuous treatment variable, *ContTreatLLM*, sums these implied changes for the focal firm over all affected LLM rivals and takes a positive value whenever *TreatLLM* takes the value one, and is zero otherwise. Thus, *ContTreatLLM* can be interpreted as the treatment dose.

Treatment across Commuting Zones—Finally, we define the subset of firm-years $\mathbf{T}_{notLLM} \subset \mathbf{T}$ that was treated by a merger however that do not observe an exogenous increase in their *CindexLLM* with any LLM competitors. Therefore, this subset of firm-years is the complement of the firm-years that experience changes in common ownership with firms within the commuting zones and thus partitions the set $\mathbf{T}$, i.e., $\mathbf{T}_{notLLM} = \mathbf{T} \setminus \mathbf{T}_{LLM}$.

The discrete (*TreatnotLLM*) and continuous (*ContTreatnotLLM*) treatment variables are then defined analogously for this subset of firms treated by mergers as their respective counterparts in the previous paragraph. In the main corpus of our paper, we use the continuous treatment variables. However, using the discrete treatment setup we obtain qualitatively consistent results.

Control Variables

Institutional ownership has been shown to influence innovation due to monitoring of managers (Aghion, Van Reenen, and Zingales, 2013, Aghion, Van Reenen, and Zingales, 2013; Guo et al., 2024, Guo et al., 2024). Since common ownership and firms' institutional ownership are related but different phenomena, we control for *InstOwn*, the percentage ownership of all institutional (13F) investors of a firm as of the calendar year-end, to disentangle both effects. We also control for *FirmSize*, which is the natural logarithm of total assets; *R&DtoAssets*, which corresponds to R&D expenses scaled by total assets; *FirmAge*, or natural logarithm of the number of years the firm has existed, according to Compustat; *PPEtoAssets* is firms' property, plant.

3.3.3 Sample and Descriptive Statistics

We combine the information from the different data sources into a firm-level panel. We start with an unbalanced sample of 8,813 unique Compustat firm identifiers and 75,402 observations. We use this large sample of firms, their pairwise ownership information, and the locations of their establishments to construct our treatment and common ownership variables. Thus, we use a comprehensive sample to include all potential LLM rivals and other portfolio firms operating in distinct labor markets to measure their common ownership with the focal firms.

Since our main outcome variables are constructed using patent information, we restrict the set of focal firms in the panel to estimate the effects of an increase in

TABLE 3.2
DESCRIPTIVE STATISTICS

Variable	*25th Perc.*	*Median*	*Mean*	*75th Perc.*	*Std. Dev.*	*N. of obs.*
Ownership:						
CindexLLM	0.001	0.002	0.003	0.004	0.003	21214
CindexALL	0.001	0.001	0.002	0.003	0.002	21214
InstOwn	0.159	0.408	0.432	0.686	0.298	21214
Firm Characteristics:						
R&DtoAssets	0.026	0.080	0.146	0.166	0.279	21214
TotalAssets (in $1M)	35.125	127.434	2074.582	621.884	10426.730	21214
FirmSize	3.559	4.848	5.089	6.433	2.123	21214
FirmAge (in years)	7.000	12.000	16.526	21.000	13.352	21214
PPEtoAssets	0.066	0.139	0.181	0.251	0.152	21214
Employees (in 1K)	0.139	0.489	5.946	2.590	22.427	21012
EmpGrowth	-0.056	0.040	0.137	0.181	1.191	19856
EmpIncrease	0.000	1.000	0.600	1.000	0.490	19856
Patent output:						
NumPatents	0.000	1.000	22.688	7.000	113.372	21214
NumAutoPatents	0.000	0.000	13.517	3.000	87.675	21214
AutoRatio	-0.693	0.000	0.039	0.693	1.264	21214
AutoDummy	0.000	0.000	0.437	1.000	0.496	21214
lnAuto	0.000	0.000	0.861	1.386	1.327	21214
lnNonAuto	0.000	0.000	0.822	1.386	1.264	21214

This Table presents descriptive statistics for our sample of patenting firms from 1990 to 2012.

common ownership to those for which we observe a positive number of patents in at least one year during our sample period in the patent data provided by Arora, Belenzon, and Sheer, 2021. The final result of our sample selection process yields an unbalanced panel of 2,006 firms, comprising 21,214 firm-year observations.

Table 3.2 reports summary statistics for the entire sample. The average *CindexLLM* is larger than the average *CindexALL*, with values of 0.003 and 0.002, respectively. This shows that the average firm has a slightly higher overlap of institutional shareholders with the average firm that operates in the same commuting zones than with the average firm operating in a disjunct set of local labor markets. Paired and unpaired *t*-tests reveal that the difference is significant at

1%, showing that common ownership of firms with labor-market overlap seems to be relevant when compared with ownership overlap in general. Institutional investors hold 43% of outstanding shares of the average firm in our sample, similar to what was found in other studies for our sample period.[10]

Moreover, the average firm in our sample invests 14.6% of total assets into research and development (R&D) activities.[11] Firms' total assets are around $2B and they are more than 16 years old, on average. The mean number of employees is approximately 6 thousand. Firms' ratio of tangible to total assets (*PPEtoAssets*) is 18.1%.

On average, firms produce around 23 patents per year of which 14 are classified as automation patents by Mann and Püttmann, 2021. However, this number hides heterogeneity across firms, and the average firm's probability to produce at least one automation patent in a year is 43.7%.

For reasons described in the following section, in our baseline estimation, we exclude observations of firms that are treated at least once during our sample period by the treatment *ContTreatnotLLM*, when estimating the effect of *ContTreatLLM*. The final set $\mathbf{T}$ of firm-years that we define as treated by a merger (as described in Section 3.3.2) consists of 1,130 firm-year observations. Of these firm-years 836 are in the set $\mathbf{T_{LLM}}$, in which the firm was affected by a merger of institutional investors that likely increases their common ownership with local labor market rivals. For a firm-year to be in the set $\mathbf{T_{notLLM}}$, we require that a firm in a particular year is affected by a merger, but that this event is not likely to increase common ownership with natural rivals in the labor market. We identify 294 firm-years in this set.

Next, we compare the average treatment doses among the two sets of treated firm-years. That is, we compare the mean dose of treatment within each treatment sample of the two subsets of firms affected by mergers with each other. For our baseline samples, the average of the 836 firms treated by mergers that likely increase common ownership with labor-market rivals is 0.022.[12]

[10] For example, Guo, Pérez-Castrillo, and Toldrà-Simats, 2019 report an ownership share of 44% belonging to institutional investors in the same years for a different sample of firms.

[11] It is well known that some firms do not report R&D expenditures in compustat. We do not replace them with zeros since this could potentially introduce errors.

[12] To help the understanding of this number, we can provide the following example that corresponds to an average treatment dose of 0.022. Assume that the focal firm has a 5% blockholder (say, investor A) that merges with another institutional investor (say, B). In the symmetric case, this other investor would hold 5% in 8.8 other firms in which A is not invested, but these firms are active in a subset of commuting zones in which the focal firm operates.

TABLE 3.3
AVERAGE CONTINUOUS TREATMENT DOSE

Set of firms:	$\mathbf{T_{LLM}}$			$\mathbf{T_{notLLM}}$			Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(1) - (4)
Variable	*Mean*	*Std. Dev.*	*N. of obs.*	*Mean*	*Std. Dev.*	*N. of obs.*	
Continuous Treatment	0.022	0.064	836	0.025	0.021	294	-0.003

This Table compares the average treatment dose within each treatment across the two treatment samples, $\mathbf{T_{LLM}}$ *and* $\mathbf{T_{notLLM}}$. The difference between the two means is has a t-score of -1.08.

The average treatment dose of the 294 firms affected by an institutional merger is larger at 0.025.[13] The average dose of treatment regarding firms outside the collection of commuting zones they operate in for firms in $\mathbf{T_{notLLM}}$ is higher compared to their counterparts in $\mathbf{T_{LLM}}$ to their labor-market rivals. However, the difference of 0.003 between the means between is not statistically significant (*t-score:* -1.08). In all our DID estimations, we report point estimates representing the economic effects of a treatment dose for the average event firm in the respective sample to facilitate interpretation and comparison.

3.3.4 Identification Strategy

We now describe in detail our identification strategy. We start by testing Hypothesis 1 through a set of two-way fixed effect regressions. These regressions, although potentially biased and suffering from endogeneity, allow us to see if we observe a general association in our panel between common ownership within local labor markets (*Cindex)* and the automation strategy of firms (*AutoRatio*) on the firm level. To this end, we estimate the model in equation (3.10).

$$\textit{AutoRatio}_{j(t+\tau)} = \beta_0 + \beta_1 CindexLLM_{jt} + \gamma X_{jt} + \alpha_j + \delta_{st} + \epsilon_{jt} \tag{3.10}$$

$\textit{AutoRatio}_{j(t+\tau)}$ is the τth lead of our main measure of the automation content of innovation as defined in Section 3.3.2. Although common owners holding shares in firms within the same local labor market may affect the investment strategy of firms contemporaneously, we expect these changes to translate into different innovation outcomes in the future, because of time lags between starting research

[13] This average treatment dose corresponds in a similar example to the merger of a 5% blockholder of the focal firm that holds 5% in 10 firms without labor-market overlap to the focal firm.

projects and the resulting patent application in case of success. Therefore, we consider $\tau = 1, 2, 3, \ldots, 6$ in the OLS panel regression. $CindexLLM_{jt}$ is the common ownership measure at the firm-year level defined in section 3.3.2. Further, we include X, the control variables discussed previously, and a set of firm and industry (s) $\times$ year (t) fixed effects to control for common shocks, e.g. industry spillovers from automation-relevant technologies or industry-specific trends in the technological feasibility frontier.

As mentioned, estimating two-way fixed effect (TWFE) models in the presence of dynamic effects may lead to biased estimates. Sun and Abraham, 2020 show that in cases such as ours, where firms are treated at different times, estimating lead or lagged models can produce biased effects, affecting model conclusions but also the researcher's ability to trust pre-trend analysis. They show that dynamic effects in TWFE models can be expressed as the linear combination of cohort-specific effects across time. For example, we cannot disentangle the contemporaneous effect of an increase in common ownership from long-term changes to the strategic direction of the firm. We therefore employ the state-of-the-art event study DID model developed in De Chaisemartin and d'Haultfoeuille, 2024 which allows us to estimate dynamic effects, under a set of more reasonable assumptions.

To derive exogenous variation in the *Cindex* we use mergers between institutional investors as in He and Huang, 2017b and Lewellen and Lowry, 2021, applying the set of continuous treatment variables defined in Section 3.3.2. As we discussed there, if institutional investors merge, they combine their portfolios, and firm-level common ownership likely increases for their portfolio firms, as we subsequently show in the data.

This identification strategy requires that financial institutions' mergers are not driven by the specific characteristics of the firms in which these institutions invest. There are several reasons why this is plausible. As He and Huang, 2017b show, about 60% of these mergers result from consolidations in the banking sector, caused by fundamental changes in the regulation of financial institutions. This led to a wave of mergers of these institutions and their asset management arms. Given the scope of the regulations and the size of the financial institutions involved, it is unlikely that the reasons behind the mergers are due to their portfolio companies. Second, Jayaraman, Khorana, and Nelling, 2002 suggest that the mergers of pure asset management institutions, i.e., the remaining 40% of the mergers, are due to strategic reasons such as exploiting economies of scale

and gaining market share. Thus, these mergers are also unrelated to portfolio firm characteristics such as innovativeness or the geographic distribution of individual firms' plants.

Having established the validity of our shock to common ownership, we describe the procedure we use to test our empirical hypotheses. Regarding Hypothsis 3, we are interested in changes of common ownership of a firm with respect to local labor market (LLM) rivals. Therefore, we use the continuous treatment variable *ContTreatLLM*, which accounts for the firms' exogenous change in common ownership with LLM rivals implied by the merger. We use this treatment variable in the estimation method presented by De Chaisemartin and d'Haultfoeuille, 2024, which is flexible to the usage of continuous treatments and our setup, in which firms may be treated several times during our sample period. As our main outcome of interest is the automation strategy of firms, we use the two automation measures, *AutoRatio* and *AutoDummy*, as the dependent variables in this experiment. Thus, we estimate the dynamic effects of treatment to LLM common ownership on the automation strategy of firms, using the universe of not(-yet) treated firms as controls. In this model, we also employ the firm characteristics described in Section 3.3.2 as control variables.

Next, regarding Hypothesis 4 in which we want to see the effects of common ownership of a focal firm concerning others, which are not natural labor market rivals of the focal firm, we apply our continuous treatment variable *ContTreatnotLLM*, analogously. Thus, we test if a firm that experiences a positive shock to common ownership, however only with regard to firms with which it does not compete for workers, increases the automation content of its innovation output in the same way, as we expect for those within local labor markets.

As mentioned, the DID method we apply accounts for the fact that firms are treated several times. However, it does not account for firms being treated by other events. Because we expect that the two treatments (increases in common ownership with regard to labor market rivals and non-rivals) are different, we exclude firms that have ever been treated by one of these treatments, when estimating the effect of the other. For instance, when estimating the effect of *ContTreatLLM*, we exclude all companies from the sample for which there is any firm year in which *ContTreatnotLLM* takes a positive value, and vice versa. For robustness, we have also estimated the effects in pooled samples. All our main results are robust to the choices regarding sample selection.

One potential criticism of our identification method comes from the critique raised in Lewellen and Lowry, 2021 that these mergers are bunched over time. There are for instance a large number of mergers around the financial crisis. If other, automation-relevant events occurred at the same time, e.g., firms increase automation to alleviate competitive pressure during the financial crisis, then our coefficients may be biased. To combat this issue we have also run the model on data up to 2006, before the onset of the crisis. All our results stay qualitatively consistent with our baseline analysis, therefore reducing this concern. We report the results in the appendix (Section C.3.2).

Another critique we address concerns the use of continuous treatment measures, instead of binary treatment variables. The key assumption we have to make is that the ownership shares as of the quarter before the announcement date of merging investors are exogenous. To address the concern that this might introduce some sort of endogeneity, we have estimated our model also using the discrete treatment variables (*TreatLLM* and *TreatnotLLM*) and report results (Section C.3.3 in the appendix) that are consistent with our estimations using continuous treatments.

Finally, we also have computed alternative measures of automation innovation based on citation counts of patents. The results reported in Section C.3.1 in the appendix are also consistent with our baseline strategy.

3.3.5 Empirical Results

In this section, we present the results of our empirical analysis.

Common Ownership of Labor Market Rivals and Automation Innovation

OLS Results—We first estimate model (3.10) on the full sample of firm-year observations. The results are shown in Table C.1 in the Appendix (Section C.2). As the results indicate, firms show a higher share of innovation output related to automation (relative to other innovations) one to five years into the future after an increase to common ownership with LLM rivals. Also, the signs of the coefficients of the control variables are sensible. Larger firms are more likely to invest in automation, probably due to economies of scale; while older firms produce relatively less automation innovation. However, as mentioned before, these results could be biased or driven by unobservable heterogeneity. Therefore, we now turn to the dynamic Difference-in-Difference model using the exogenous

FIGURE 3.2
DYNAMIC EFFECTS: COMMON OWNERSHIP

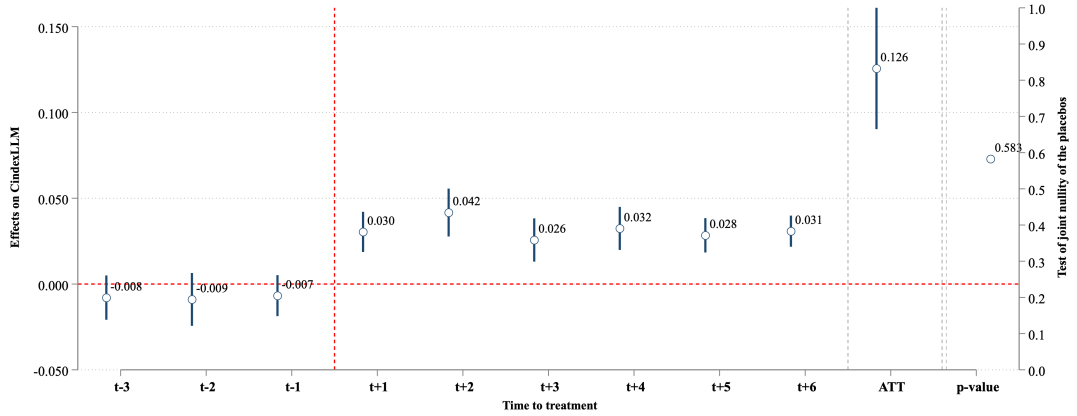


Figure 3.2: *The dynamic effect of being treated by an institutional merger that likely increases common ownership with LLM rivals (ContTreatLLM) on the raw firm-level common ownership regarding LLM rivals (CindexLLM).*

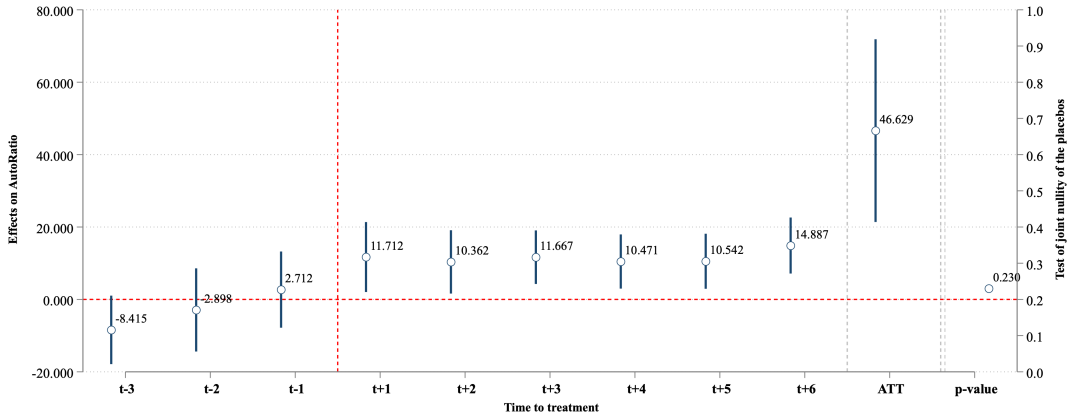
changes in common ownership.

Relevance of Treatment—To test Hypothesis 3 using the DID model, we first show that mergers between institutional investors lead to an increase in the average common ownership with LLM rivals, using the continuous treatment *ContTreatLLM* and the *CindexLLM* as the dependent variable. Lewellen and Lowry, 2021 show that institutional mergers lead to an increase of common ownership on the firm-pair level. One concern may be that after merging, the merging institutional investors or other institutions adjust their portfolios such that the effect on common ownership could be negligible, or disappear quickly. Also, it is crucial in our setup that an increase in common ownership with an LLM rival on the firm-pair level is not compensated by other changes in common ownership with other LLM rivals, since our outcome variables are defined on the firm level.

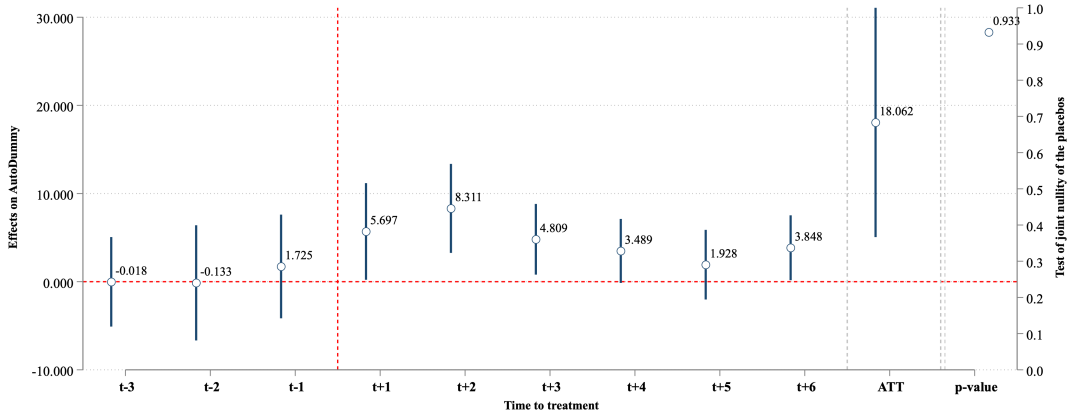
Figure 3.2 indeed shows a jump in the average *CindexLLM* following treatment in year 0. The merger event leads to an increase of around 0.03 percentage points in firm level common ownership per year. This effect is persistent during the six years following treatment.[14] The pre-treatment period effects are not significant and the p-value of joint nullity of the placebos is 0.583, which indicates that the parallel-trends assumption is satisfied. Therefore, we can conclude that treated firms experience a common ownership increase with other firms in the same

[14]As defined in Section 3.3.2, the dependent variable is the average ownership overlap of the focal firm with all its labor market rivals.

FIGURE 3.3
DYNAMIC EFFECTS: AUTOMATION MEASURE



(a) Using the continuous measure *AutoRatio*



(b) Using the binary measure *AutoDummy*

Figure 3.3: *This Figure shows the dynamic effects of within-LLM increases to common ownership (ContTreatLLM) on the automation strategy of firms.*

local labor markets.

Automation Strategy—We now turn to our main outcome variables to test Hypothesis 3. The principal model uses *AutoRatio* as the dependent variable to test whether the firm changes its innovation strategy to become more automation-focused, controlling for a potential change in total innovation. We also use the automation indicator variable, *AutoDummy* to test if the exogenous changes to common ownership affect firms' propensity to invest in automation.

Figure 3.3 shows that an increase in common ownership with local labor market rivals due to a merger leads to a significant increase in the automation content of innovation for the treated firms in $\mathbf{T_{LLM}}$ over the six years following treatment.

This increase is also economically significant, as it corresponds to a change of 46.6% for the average treated firm, over the six year period. In the discrete treatment setup, the resulting change in the ratio of automation to non-automation patents reported in the Appendix (Section C.3.3) is 49.6%. On a yearly basis, we see that the automation content of innovation increases significantly by around 10% year on year.

The propensity of firms to produce automation patents significantly increases for each of the first three years after the average merger event (between 4.8 and 8.3 percentage points), as indicated in Panel (b) of Figure 3.3. After that, the effect remains positive but becomes insignificant. Over the entire 6 year period we see an 18.1% increase in the probability of patenting an automation innovation over the unconditional mean. Again for both the continuous and dummy automation outcome measure the parallel trends conditions are satisfied.

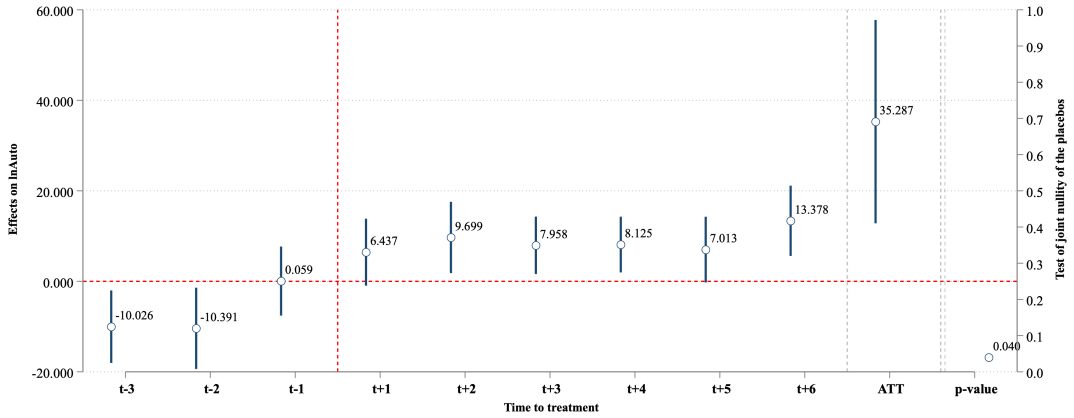
Next, we use individual patent count variables, *lnAuto* and *lnNonAuto* to examine separately which type of patenting drives the effect estimated. The indicator variable *AutoDummy* suggests that more firms are likely to patent for automation innovations upon treatment. However, we want to confirm whether, on the intensive margin, the automation content of innovation increases due to more automation patents and not because of a decrease in non-automation patents.

The results are shown in Figure 3.4. Treated firms experience a surge in automation patent output. The ATT is highly significant and indicates an increase of 35.3%. In years two to four, the significant increase in the number of automation patents is between 7.9% and 9.7%. On the contrary, the ATT as well as the yearly effects on non-automation patents are not statistically different from zero. The ATT for non-automation patenting is in fact negative, pointing to a potential strategy shift of firms from non-automation to automation. The pre-trend results for the automation patents however show a potential anticipatory effect, where there is a pre-event increase in automation patenting, this requires further examination.

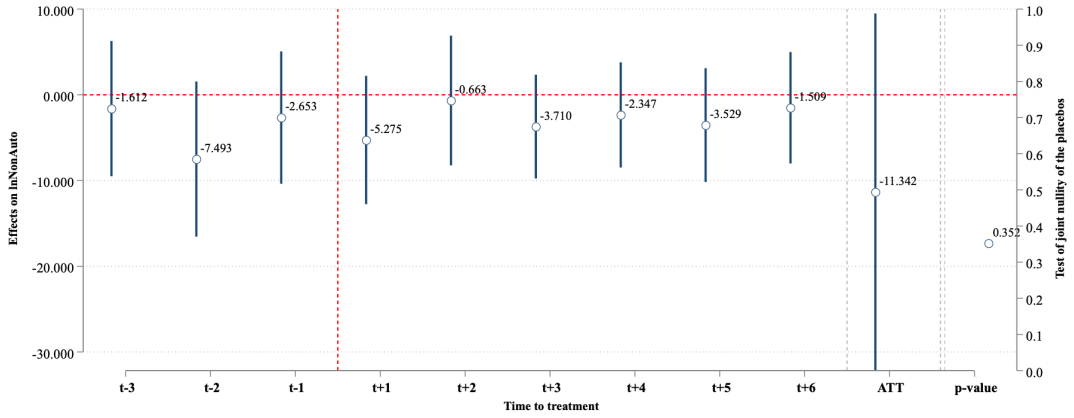
The results shown in this section are robust to using the discrete treatment, as well as using patent citations to compute the automation strategy measures. Furthermore, we have also estimated the effects for our sample until 2006. The results are qualitatively similar. The robustness checks can be found in Section C.3 of the Appendix.

Overall, our results suggest that increases in common ownership between firms

FIGURE 3.4
DYNAMIC EFFECTS: AUTOMATION VERSUS NON-AUTOMATION



(a) Effect on automation patents



(b) Effect on non-automation patents

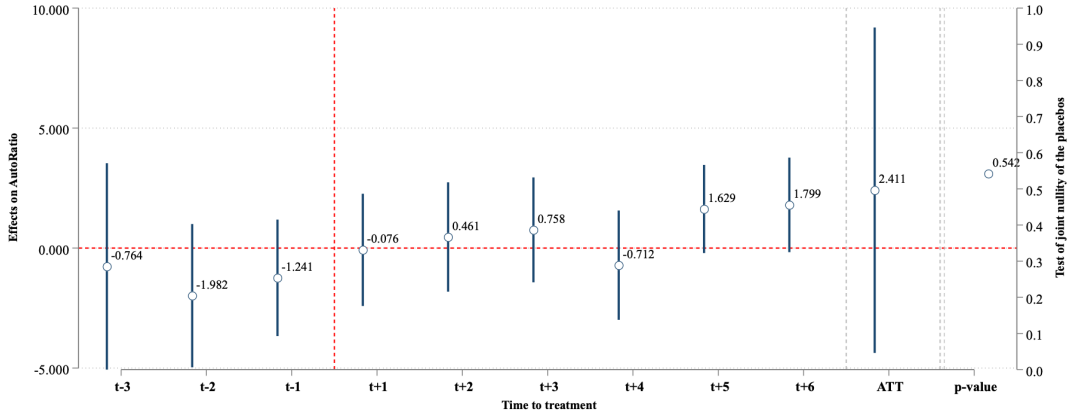
Figure 3.4: *This Figure shows the dynamic effects of within-LLM increases to common ownership (ContTreatLLM) on the automation and non-automation patents.*

that compete for workers increase their automation-related innovation output. In the next section, we will use our alternative treatment to study the effect of common ownership on automation when labor market rivalry is absent.

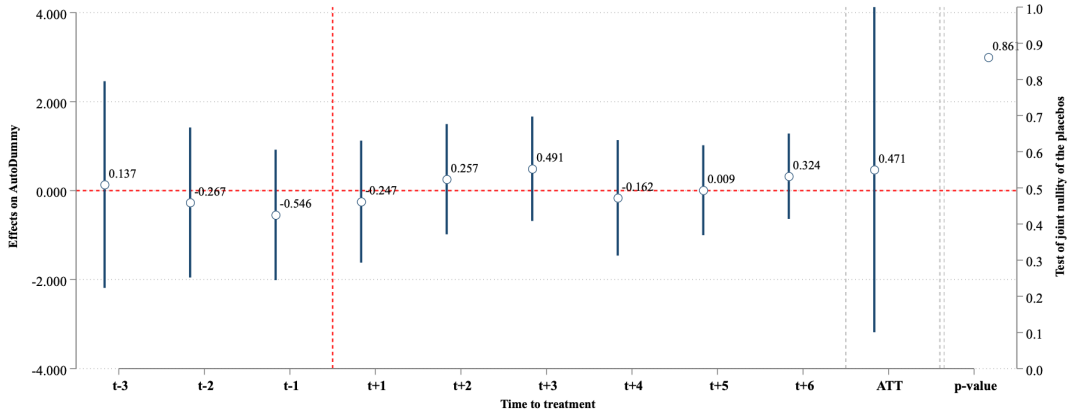
Common Ownership and Automation Innovation in the Absence of Labor Market Rivalry

We now study how increases in common ownership affect firms' automation strategy in the absence of labor market rivalry. That is, we focus on the change in automation innovation of focal firms that experience an increase in common ownership with other firms outside the commuting zones in which the focal firms operate. Thus, we are testing Hypothesis 4.

FIGURE 3.5
DYNAMIC EFFECTS: NO LLM OVERLAP



(a) Using the continuous measure *AutoRatio*



(b) Using the binary measure *AutoDummy*

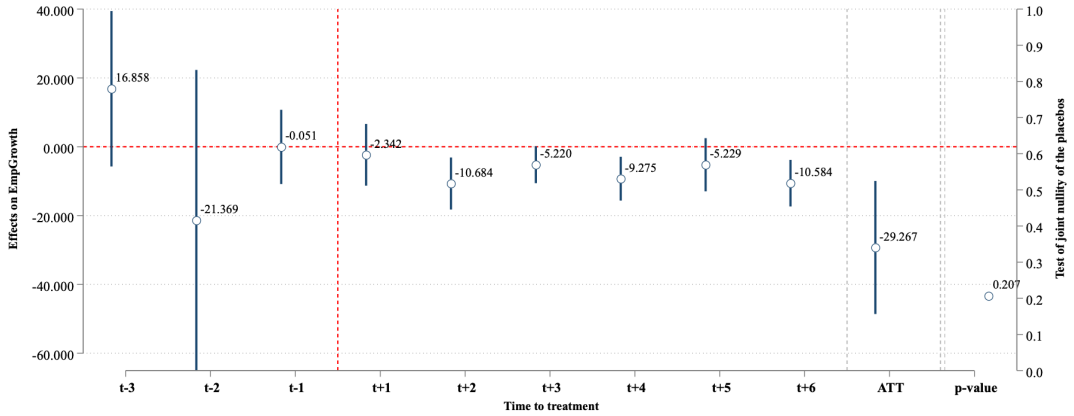
Figure 3.5: *This Figure shows the dynamic effects of across-LLM increases to common ownership (ContTreatnotLLM) on the automation strategy of firms.*

The results are shown in Figure 3.5. The ATTs for both, the continuous *AutoRatio* and the binary *AutoDummy* are insignificant from zero. Furthermore, all of the yearly effects are not statistically different from zero. Also, in terms of magnitude, while they are positive, the effects are much smaller than for those firms experiencing increases in common ownership with LLM rivals. The results therefore corroborate Hypothesis 4.

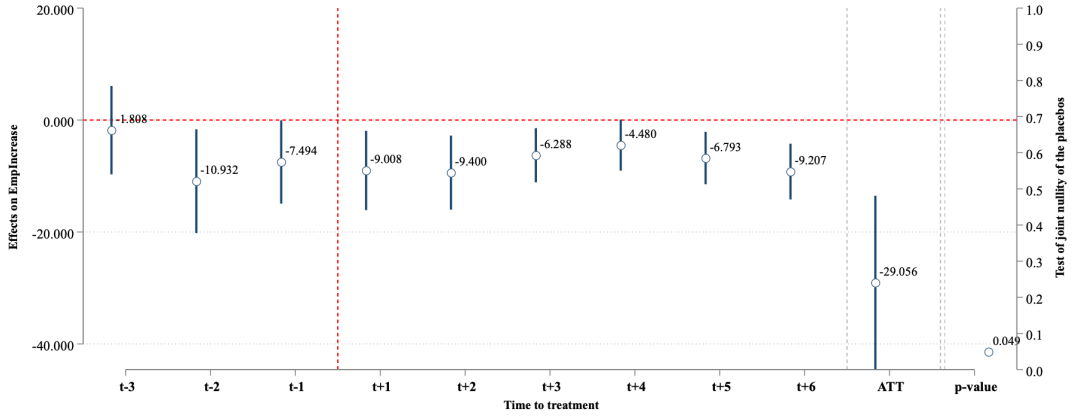
The Employment Effects of Common Ownership under Labor Market Rivalry

We have shown that increases in firms' common ownership within local labor markets affect firms' automation strategy. Increases in within-LLM com-

FIGURE 3.6
DYNAMIC EFFECTS: EMPLOYMENT



(a) Using the continuous measure *EmpGrowth*



(b) Using the binary measure *EmpIncrease*

Figure 3.6: *This Figure shows the dynamic effects of within-LLM increases to common ownership (ContTreatLLM) on firms' growth in terms of employment.*

mon ownership raise the automation content of firms' innovation output. We now turn to the outcomes in terms of employment. Using our treatment *ContTreatLLM*, we now estimate the effect of common ownership within labor markets on firms' hiring decisions, applying *EmpGrowth* and *EmpIncrease* as the outcome variables. The results are shown in Figure 3.6.

In the six years after the treatment event firms' employment growth rates and their likelihood of having positive employment growth both decrease significantly. On a yearly basis, their growth rates decrease up to 10.7 percentage points. Their probability of experiencing positive growth rates in employment decreases by 4.4 to 9.4 percentage points. The pre-trends are weaker and nosier than the main results and therefore these results need further examination.

3.4 Conclusion

We develop and test a theory of the impact of common ownership on firms' automation strategies. We show, both theoretically and empirically, that increases in common ownership of firms with local labor market rivals lead to increases in the number of automation patents and the overall automation content of the firms' innovation output. We measure automation using the texts for patents produced at each firm.

Thus, we provide evidence that institutional common ownership influences firms' innovation strategy and the direction of technological change, steering portfolio firms to focus more on automation in their innovation process. Moreover, we do not find evidence that exogenous changes in common ownership of a focal firm with those companies that operate in distinct labor markets cause an increase in automation. This result is consistent with the mechanism we developed in our model. That is, institutional common ownership increases firms' incentives to automate to reduce labor market competition among portfolio firms. Consistent with this mechanism, we observe that increases in firms' common ownership with labor market competitors reduce firms' future employment growth.

The implications of these results are critical for policymakers concerned with the effects of technological change, especially the advancement of automation technologies, on social welfare and inequality.[15] Our research demonstrates that the substantial rise in common ownership, observed in both the US and Europe, further incentivizes firms to develop and implement technologies aimed at substituting human labor.

[15] See Acemoglu and Restrepo, 2018, Moll, Rachel, and Restrepo, 2022, Santini, 2024.

A. Chapter 1 Appendix

A.1 Inferring the Knowledge Space: Intuition

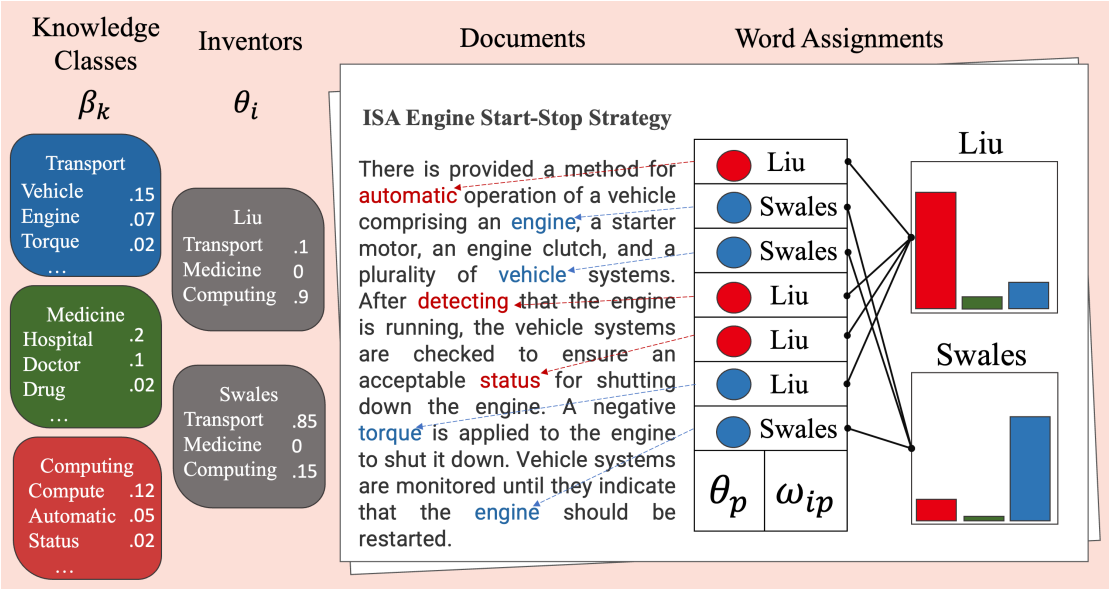
Figure A.1 provides an intuitive example of how the Author-Topic model learns the knowledge classes, inventor knowledge profiles and can describe the knowledge content of a patent and an inventor's contribution share. The knowledge classes (β_k) and inventor knowledge distributions (θ_i) are latent variables. The idea being that for each word, in every patent, the model attaches a knowledge class $\times$ inventor pair, thus revealing the latent variables over many patent documents.

Over many patents, if the words *vehicle*, *engine*, and *torque* appear together, they start to form the transport class. The same goes for the words *hospital*, *doctor* and *drug*. In the example given, the word automatic was contributed by Liu using the computing class, the word engine however was contributed by Swales using the transport class. If an inventor appears on patents using words from one of these knowledge classes, their weight on that knowledge class increases. This is the basic functionality of the author-topic model as written in the *Gensim* package. However, more information can be extracted. Naturally, as each word is matched with an inventor $\times$ knowledge class pair, you can describe the knowledge content of the patent as a mixture of knowledge classes. Most important for this chapter, the number of words reveals the contribution of each inventor.

As written here, this inference method matches the Gibbs Sampling algorithm. In this paper I employ a Variational Bayes method. The Gibbs Sampling algorithm would allow you to make the same calculations as in this example, and sum words across inventors, knowledge classes within a patent. The Variational Bayes instead uses an approximation to this method, which over large patent data is more efficient. The Gibbs sampler converges on the true solution, while the variational Bayes inference method converges on an approximation to the true solution.

FIGURE A.1

INTUITIVE LDA EXAMPLE



Notes: An intuitive example of how LDA works. The example used is a paraphrased version of USPTO patent number US6752741 which expires 2022-05-31. The knowledge class and inventor parameters are learnt by iterating over patent texts and allocating inventors and topics to words.

A.2 Additional Tables and Figures

TABLE A.1

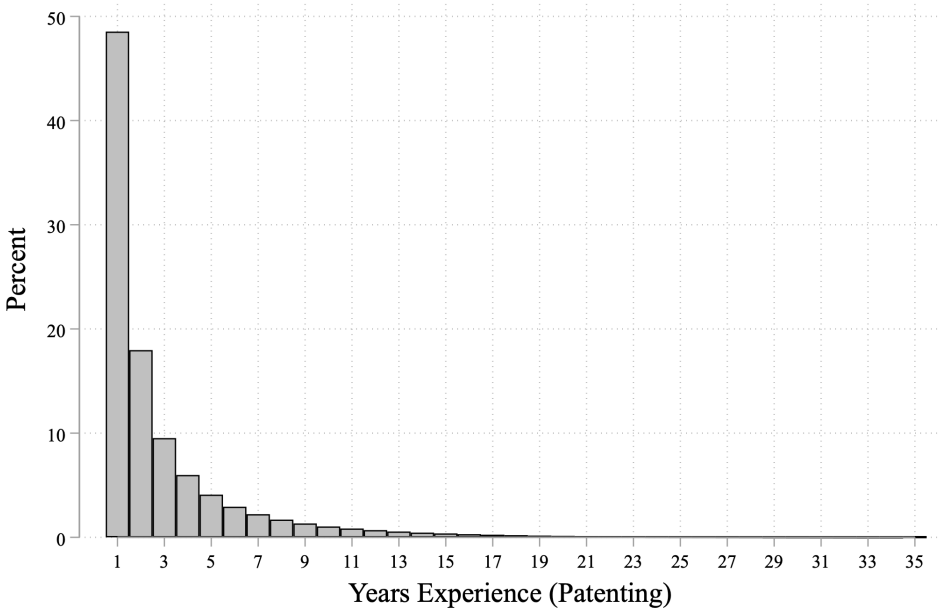
DESCRIPTIVE STATISTICS

	Mean	Std. Dev.	Min	Max
		Patent Level		
Citations	16.238	45.038	0.000	1896.000
Market Value (m)	3.699	19.526	0.000	1341.500
Pr(Breakthrough)	0.130	0.336	0.000	1.000
New vocabulary reused	15.909	89.962	0.000	6878.000
New vocabulary	3.663	17.341	0.000	1028.000
Year	2002.918	5.761	1991.000	2021.000
Observations	29332			
		Inventor Level		
No. Patents	5.266	7.802	1.000	399.000
No. Teams	1.783	2.262	1.000	104.000
Female	0.080	0.271	0.000	1.000
Observations	34613			
		Team Level		
Team Size	6.171	3.106	2.000	51.000
No. Patents	3.702	11.535	0.000	153.000
Concentration	15.661	9.421	0.035	112.328
Observations	10000			

Notes: The table reports the mean, standard deviation, minimum, and maximum values. Observations represent the number of patents, inventors, or teams in each category. This is the sample used in the main analysis, not to train the LDA. The descriptive statistics for the full sample of 1.2 million patents is given in the appendix.

FIGURE A.2

HISTOGRAM OF INVENTOR EXPERIENCE



Notes: A histogram of inventor total years experience. Each inventor is included once, where the total number of years experience corresponds to the number of individual years in which they appear on a patent. This is taken over the full sample of USPTO data from 1976 to 2024.

A.3 Gender Results

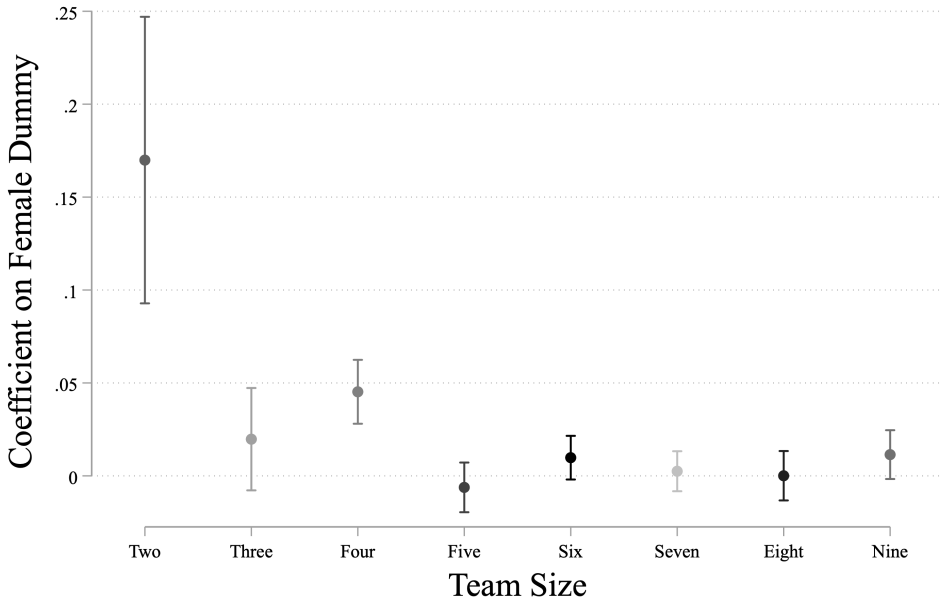
I demonstrate how individual characteristics explain the contribution share of an inventor to a team project. I define a gender diverse team as one that contains at least one man and one women. I define a female dummy which is equal to one if the inventor is a woman. I then introduce a team size dummy δ_m and control for a year fixed effect through δ_t.

I measure experience as the cumulative count of the number of patents that inventor has produced, prior to patent p. I split this variable into 4 equally sized bins to track the experience of inventor i on patent p ($exp_{i(p)}$). These refer to low, medium-low, medium-high and high experience levels. I build a second count for the mean number of patents the inventor's collaborators in team τ have produced, prior to the patent p. I denote the team τ minus inventor i by $\tilde{\tau}$. I split this same variable into the same four bins ($exp_{\tilde{\tau}(p)}$). This tracks whether inventor i collaborated with junior or senior co-inventors.

$$\omega_{ip} = \beta_0 + \beta_f \text{female}_i + \sum_{e=1}^{4} \beta_e \cdot \mathbb{1}(\text{exp}_{i(p)} = e) + \delta_m + \delta_c + \delta_t + \epsilon_{ip} \tag{A.1}$$

FIGURE A.3

CONTRIBUTION SHARE OVER GENDER



Notes: This figure presents estimated coefficients on the female dummy from regressions of inventor contribution share on gender, controlling for year, team size, technology class, inventor experience, and patent impact. Separate regressions are estimated for teams of sizes 2 to 9, conditional on gender diversity (i.e., teams with both male and female inventors). Confidence intervals at 95% level are shown.

For this subset of teams, I find that women contribute more on smaller teams. The coefficient is only significant for teams of two and four. However, this difference disappears entirely for larger teams.

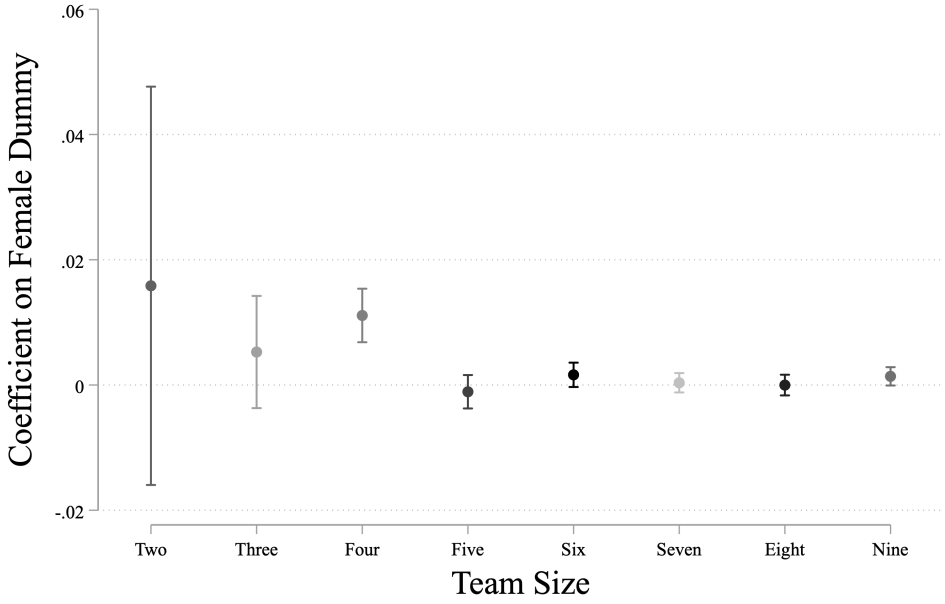
$$\begin{aligned}\omega_{ip} = \beta_0 + \beta_f \text{female}_i + \sum_{e=1}^{4} \beta_e \cdot \mathbb{1}(\text{exp}_{i(p)} = e) + \sum_{o=1}^{4} \beta_o \cdot \mathbb{1}(\text{exp}_{\tilde{\tau}(p)} = o) \\ + \sum_{e=1}^{4}\sum_{o=1}^{4} \delta_{eo} \cdot \mathbb{1}(\text{exp}_{i(p)} = e) \times \mathbb{1}(\text{exp}_{\tilde{\tau}(p)} = o) + \delta_m + \delta_c + \delta_t + \epsilon_{ip}\end{aligned} \tag{A.2}$$

This effect however is driven by the relative experience of each inventor to the team. Figure A.4 shows the results for when I introduce the same interaction

effect in the main body to equation A.1. I find that the contribution gap is much smaller, and now statistically insignificant from zero. This means that the fact that women contribute more than men on gender diverse teams can be explained by the fact that they tend to be relatively more junior colleagues. Therefore the conclusions presented hold given their relative inexperience, not a feature of their gender.

FIGURE A.4

CONTRIBUTION SHARE OVER GENDER | TEAM EXPERIENCE



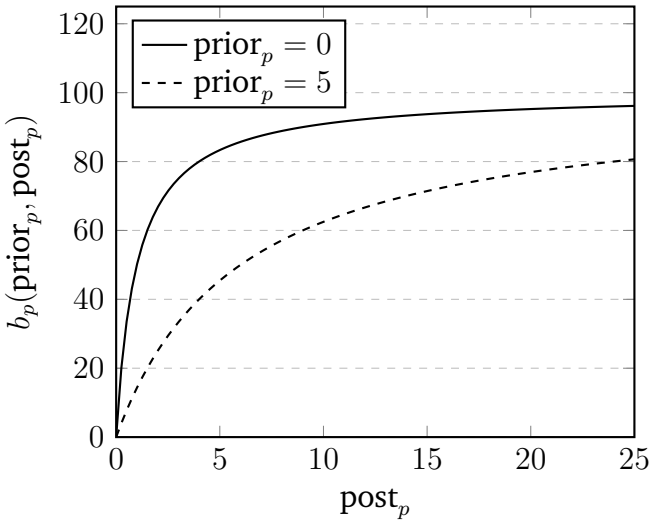
Notes: This figure presents estimated coefficients on the female dummy from regressions of inventor contribution share on gender, controlling for year, team size, technology class, inventor experience, and patent impact. This regression model introduces an interaction between the inventor's experience level and that of their co-inventors. Separate regressions are estimated for teams of sizes 2 to 9, conditional on gender diversity (i.e., teams with both male and female inventors). Confidence intervals at 95% level are shown.

B. Chapter 2 Appendix

B.1 Additional Tables and Figures

FIGURE B.1

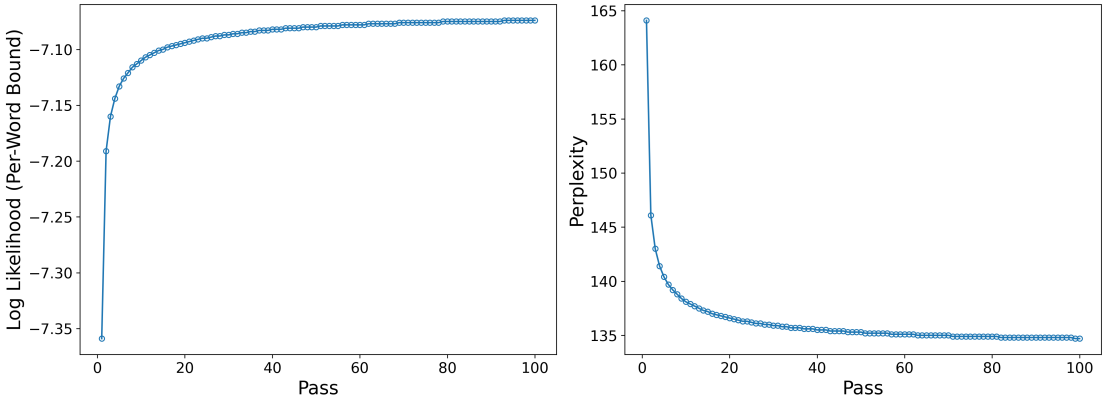
RAW BREAKTHROUGH MEASURE



Notes: Function for the raw breakthrough measure at the patent level plot over a generated range of post-count values. This measure is bounded between 0 and 1, but importantly captures a concept of percentage change even when the pre-count is equal to zero.

FIGURE B.2

LDA MODEL CONVERGENCE

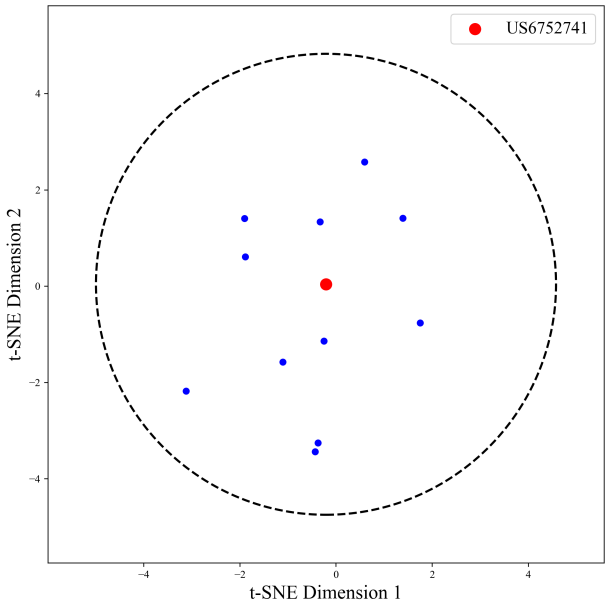


Notes: Two convergence plots. One showing the log likelihood per word bound, and the second the model perplexity. Each statistic is calculated after every 25 iterations over the data.

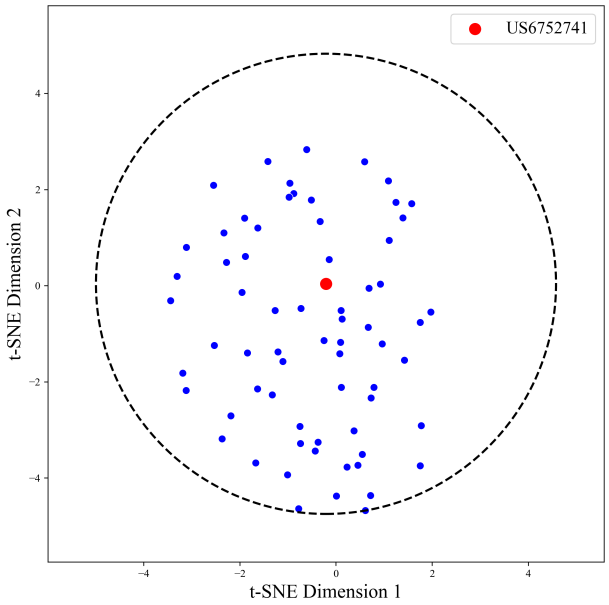
FIGURE B.3

VISUALISING THE 50-DIMENSIONAL PATENT RESEARCH FIELDS

U.S. PATENT 6752741: PRE PUBLICATION



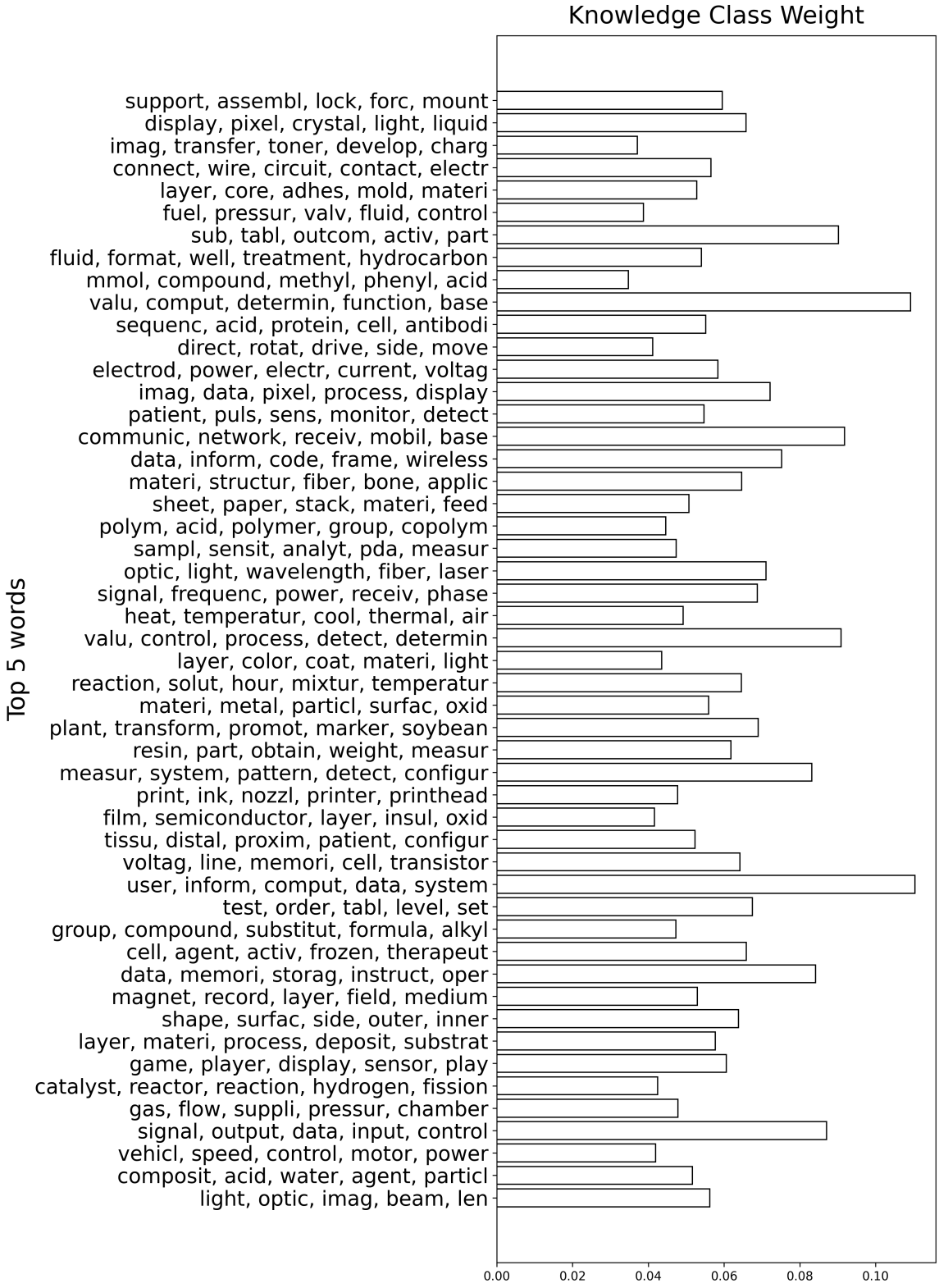
U.S. PATENT 6752741: IN 2019



Notes: Pre & Post Publication for US6752741 titled " ISA Engine Start Stop Strategy". This figure visualizes the t-SNE embeddings of patent topic distributions for the selected target patent. The 50-dimensional patent embedding θ_p is reduced to two dimensions using t-SNE, a dimensionality reduction technique optimized for capturing relative similarities between points in lower-dimensional space. In each panel, the red marker highlights the target patent, while other blue markers represent additional patents. The top panel shows only patents published prior to or in the same year as the target patent, while the bottom panel includes the full sample. The dashed black circle, centred on the target patent plots an example research field for a given radius r.

FIGURE B.4

INFERRED BAYESIAN PRIOR α



Notes: Learnt α Dirichlet prior from the `Gensim` package option auto. The Y-axis presents the 5 words with the largest weight within the knowledge class to word distribution for that class. The height of the bar represents the weight on that class in the Bayesian prior.

FIGURE B.5

AGGREGATE TOPIC DISTRIBUTION BY PATENT TYPE

Notes: Plots the average knowledge class distribution by patent type. The data on Green, Automation and Cancer patents provided by PatentsView (2024), Mann and Püttmann (2023), and Cancer Moonshot: USPTO (2024). Again the top 5 words per class shown on the Y-axis.

FIGURE B.6

TEAM STATISTICS

Notes: Four binned scatter plots produced with `binscatter`*. Each plot is taken for the average within one year, across all teams who first patented in that year. The total number of patents is defined in equation 2.8. Volume is defined in equation 2.9 as the square root of team size multiplied by the weighted average of the maximum euclidean distance between team member knowledge points, and the mean distance. I refer to this weighted average as the team breadth. The bottom two panels represent both parts of the multiplicative volume measure defined.*

TABLE B.1

PATENT REGRESSION ESTIMATES: DIRECTION

	Dependent variable: Pr(Direction)			
Prior work | Direction$_{pt}$	0.0290***	0.0281***	0.0281***	0.0438***
	(41.51)	(44.80)	(44.88)	(47.19)
Prior work | Direction$_{pt}$ Sq.				-0.0001***
				(-24.85)
N	1218385	1218385	1218366	1218366
Controls	✓	✓	✓	✓
Direction FE	✓	✓	✓	✓
Year × Direction FE		✓	✓	✓
Team size			✓	✓

Notes: Each column corresponds to a logistic regression of the probability a patent is one of three types (z), where all three types are stacked into one regression model. The dependent variable is composed of three binary indicators for whether that patent achieves each of the three directions: mitigates climate change, reduces cancer risk or automates production. All standard errors are clustered at the knowledge cluster × year level. Controls include $d(\theta_p^e, \theta_p)$.

TABLE B.2

PATENT REGRESSION ESTIMATES: BREAKTHROUGH

	Dependent variable: Pr(Breakthrough)			
Prior work$_{pt}$	-0.0019***	-0.0014**	-0.0014**	0.0008
	(-3.76)	(-2.88)	(-2.89)	(1.60)
Prior work$_{pt}$ Sq.				-0.0000**
				(-2.97)
N	408772	408772	408753	408753
Controls	✓	✓	✓	✓
Year FE		✓	✓	✓
Team size			✓	✓

Notes: Each column corresponds to a logistic regression of the probability a patent is either a breakthrough. The dependent variable is the probability that patent is in the top 75% of the breakthrough score. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). All standard errors are clustered at the knowledge cluster × year level. Controls include $d(\theta_p^e, \theta_p)$.

TABLE B.3

TEAM TREATMENT ESTIMATES: HETEROGENEOUS

	Dependent variable: Pr(Breakthrough)		
	Low	Medium	High
$D_{\tau t}$	-0.0051*	-0.0023	0.0040*
	(-2.15)	(-1.20)	(2.27)
Prior work$_{\tau_1 t}$	-0.1972***	-0.0586***	-0.0088***
	(-4.67)	(-4.78)	(-3.59)
Volume$_{\tau}$	-1.9697*	-4.6353***	-1.1707
	(-2.41)	(-4.89)	(-1.22)
N	3011	2770	2425
Controls	✓	✓	✓
Team FE	✓	✓	✓
Period FE	✓	✓	✓
Year FE	✓	✓	✓

Notes: All regressions are team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable for panel I) is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

Table B.3 shows the coefficient for the regression model, split over each of the three terciles of the prior patent count in a team's knowledge field. The first row coefficients are shown in Figure 2.10. The second row coefficients on Prior work$_{\tau_1 t}$ require further explanation. This figure is a within-team model, since all regressions include a team fixed effect. I argue that this model demonstrates that teams in the lowest tercile are in fact on the upward sloping region of the inverted-U shape shown in Figure 2.9. This may seem at odds with the fact that the coefficient on Prior work$_{\tau_1 t}$ is negative for all three sub-samples. First of all, due to the non-linear nature of the logit model, the size of these three coefficients should not be compared across models. Since as average Prior work$_{\tau_1 t}$ increases across each sample, this drives the coefficient towards zero. Second,

when the model is run across each sub-sample on a logit model using between team variation, then the coefficient on Prior work$_{\tau_1 t}$ for the lowest tercile is positive, and for the largest tercile is negative. This is inline with the inverted-U hypothesis. However this does suggest that there are important differences for between versus within team changes that deserve further exploration.

TABLE B.4

TEAM TREATMENT ESTIMATES: NOVEL PATENTS

	Low	Medium	High
	1.	2.	3.
Words	0.00001	-0.00004	0.00008**
	(0.25)	(-0.38)	(3.11)
Bi-Grams	0.00002	-0.00007	0.00015***
	(0.17)	(-0.36)	(3.35)
Tri-Grams	0.00005	-0.00040	0.00016**
	(0.30)	(-1.66)	(2.99)
N	9813	10041	10588
Controls	✓	✓	✓
Team FE	✓	✓	✓
Period FE	✓	✓	✓
Year FE	✓	✓	✓

Notes: All regressions are team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable for each of the three models is taken from Arts, Hou, and Gomez, 2021 and counts the number of new n-grams introduced by each patent. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

B.2 Contribution Weights Validation

This chapter utilises the new method of inferring the contribution of each team member to the knowledge contained in a patent developed in chapter 1. To demonstrate the power of this method, I validate the inventor contribution weights using a prediction model. I propose that if the weights capture information on the true contribution share of each inventor, then the patenting history of inventors who contribute significantly more should be a stronger determinant of the technology classification awarded to a patent. This is an application of the validation method introduced in Chapter 1.

TABLE B.5
VALIDATION OF THE CONTRIBUTION WEIGHTS

	$\%\Delta \geq p90$		$\%\Delta \geq p75$		$\%\Delta \leq p25$		$\%\Delta \leq p10$	
T-Test	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Lead	57.126	0.019	56.806	0.015	50.244	0.045	50.030	0.031
Second	42.874	0.019	43.194	0.015	49.755	0.045	49.970	0.031
Difference	14.251	0.027	13.612	0.021	0.488	0.064	0.059	0.043

Notes: T-test to determine differences across lead and second inventor feature importance. Small and large gaps are defined by the percentiles on the percentage difference $\%\Delta$ between the lead and second inventor. After each of the 50 runs of a random forest I calculate the total feature importance for the lead and second inventor patent histories. The features are the top five most common CPC classes used by the lead, and the second inventor. The target variable is the CPC class awarded to the patent. The final T-test is calculated over N=50.

I propose that if the gap between the contribution shares of the two inventors is large, then the lead inventor's patenting history will be a significantly stronger predictor of the CPC class awarded to a patent. While if that difference is small (both inventors contributed similarly to the patent), then I predict there to be no significant difference. This corresponds to the total feature importance for the lead inventor's patenting history being significantly larger than that of the second inventor. The table shows that for teams in which the difference between the first and second leader's contributions is large (top 90%) then the lead inventor's history provides on average 14.251% more information when predicting the focal patent's technology class. Whereas for teams in which the difference is small (bottom 10%), this difference disappears.

B.3 Hypothesis Development

When a team τ draws contribution shares ω_p to define their expected patent knowledge distribution θ_p^e within local knowledge field $B(\theta_p^e, r)$. A local knowledge field and time define a breakthrough score (b_p) and innovation direction (z_p) tuple

$$(b_p, z_p) \mid \theta_p^e, t.$$

The breakthrough score measures the scientific impact of that innovation. Did it spark a new and successful research field? The direction of an innovation measures the target use of the patent. Does that innovation achieve a certain goal, for example to mitigate climate change, reduce cancer risks or automate production?

The idea being that the impact of an idea is time dependent. The most straightforward example is that there is a significant gain in being the first to invent a new object. If you are working on an artificial intelligence innovations, the same idea has a different value today than it would have had fifty years ago, when many AI models were first theorised. In terms of being a breakthrough, there are now plenty of AI patents which have come before. But the direction—the ability of this combination to meet a specific objective—depends on whether similar innovations have previously achieved that goal. If past efforts with similar knowledge combinations have achieved certain outcomes, similar innovations may continue along that path, shaping the future of innovation in that area. Timing plays a critical role, as the same combination might be more or less effective depending on the state of knowledge and technological demand at the time. To complete the prior example, inventors have a wealth of prior AI knowledge to use when automating production today when compared to the past.

Both b_p and z_p are modelled as latent variables, such that for both $y_p \in \{b_p, z_p\}$

$$y_p(\theta_p^e, t) = \begin{cases} 1 & \text{if } f_y(\theta_p^e, t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.1}$$

Allowing for an abuse of notation, f_y is a general function that maps a team's location in knowledge space to the real line. This function can be mapped into the probability that a given patent achieves that outcome. I will now link this

set-up to both hypotheses outlined in section 2.2. A similar argument can be easily derived for whether a patent targets a given direction, as discussed in section 2.6.1.

Hypothesis 1 proposes that there exists an inverted-U shape relationship between the probability a patent is a breakthrough and the quantity of prior work on which it builds. Formally this is given by

$$\frac{\partial f_b}{\partial n_{pt}} > 0 \quad \text{and} \quad \frac{\partial^2 f_b}{\partial n_{pt}^2} < 0.$$

This can be tested directly in the data through a logit model that captures the latent variable structure. Given the definition of a team span in equation 2.4, we can define the expected value for each outcome as the following.

$$\mathbb{E}[y_p|\tau, t] = \frac{1}{\text{vol}(S(\tau))} \int_{S(\tau)} y_p(\theta_p^e, t)\, d\theta_p^e \tag{B.2}$$

In other words, what proportion of all the teams potential ideas achieve outcome y? Each potential project can be given a probability of being a breakthrough or not, through the latent variable model outlined. Therefore since all projects are drawn with a uniform probability, I can test the expected team patent outcome again using a logit model. Thanks to the uniform distribution assumption, the expected value defined in equation B.2 relies on the volume of a team span. Therefore when testing hypothesis 2 I use a change in the density of patents within a team's field to distinguish each case.

B.4 Counting Objects in Knowledge Space

To build count of the number of patents p' within the local knowledge field of a target patent p, it is straightforward to find all patents such that $\rho(\theta_p, \theta_{p'}) \leq r$. A patent p belongs to team span $S(\tau)$ if there exist a set of weakly positive weights that sum to 1 across the team member distributions to form a convex combination equal to the distribution for that patent.

To solve whether a patent p belongs to the local knowledge field of a team of n_τ members, I first find the closest point $\tilde{\theta} \in S(\tau)$ to that patent by finding the solution to the following problem.

$$\min_{\omega \in \mathbb{R}^n_\tau} \left\| \theta_p - \sum_{i \in \tau} \omega_i \theta_i \right\|$$
$$\sum_{i \in \tau} \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0$$

The objective is too choose the set of weights, such that they form a convex combination of each team members knowledge distribution, to minimise the distance between that point and the target patent distribution. If the distance between these two points is zero then this patent belongs to the convex hull of the team. If this distance is below the defined radius r, which remains constant across patents and teams, then this patent belongs to that teams local knowledge field.

I need to solve this problem for all patents in the sample, for each team. This is a huge number of problems to solve, in order to reduce the computational burden I take the following mathematical short cut. I first calculate the centroid of the team span $S(\tau)$ as

$$c = \frac{1}{n_\tau} \sum_{i \in \tau} \theta_i$$

Calculate the maximum distance from the centroid to any point within the team vector using

$$d_{\text{max}} = \max \|\theta - c\|$$

using the euclidean norm. For each patent θ_p calculate the distance between that patent distribution and the centroid $d = \|\theta_p - c\|$.

Notice that any point which is further form the centroid than the maximum distance within the team span plus the radius r cannot form part of the local

knowledge field. Therefore only solve the problem specified for those patents which

$$d_i \leq d_{\text{max}} + r$$

Since this calculation is computationally far less demanding and faster than solving the problem, but ultimately gives the same solution.

B.5 Robustness Tests

TABLE B.6
BREAKTHROUGH RESULTS: NO REPLACE SAMPLE

Dependent variable: Pr(Breakthrough)				
	1.	2.	3.	4.
$D_{\tau t}$	-0.0005	0.0043***	0.0031*	0.0015
	(-0.70)	(3.42)	(2.53)	(1.25)
Prior work$_{\tau_1 t}$	-0.0007***	-0.0271***	-0.0166***	-0.0165***
	(-9.95)	(-5.29)	(-4.44)	(-4.40)
Volume$_{\tau}$				-1.5727
				(-1.55)
N	25014	7758	7758	7758
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

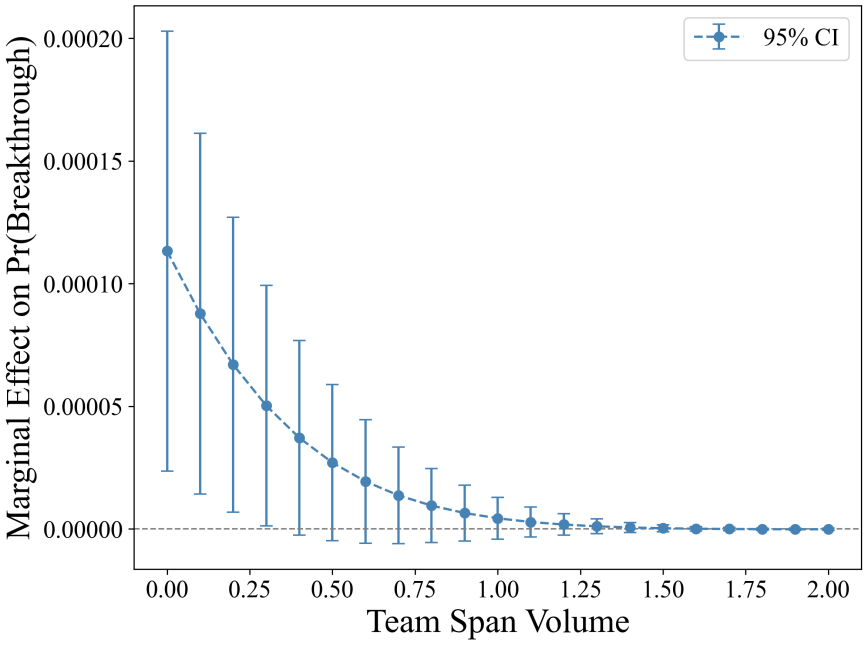
Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable is an indicator for whether the patent is a breakthrough using the Kelly et al., 2021 data. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee. The sample is a subset of the full sample, including all control teams, and teams treated by the death of a team member who did not replace the deceased team member.

TABLE B.7
BREAKTHROUGH RESULTS: REPLACE SAMPLE

	Dependent variable: Pr(Breakthrough)			
	1.	2.	3.	4.
$D_{\tau t}$	0.0006***	0.0053***	0.0038**	0.0027*
	(3.72)	(3.47)	(3.13)	(2.31)
Prior work$_{\tau_1 t}$	-0.0008***	-0.0208***	-0.0120***	-0.0125***
	(-11.36)	(-6.03)	(-5.16)	(-4.96)
Volume$_{\tau}$				-2.8226***
				(-4.12)
N	29655	9325	9325	9325
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable is an indicator for whether the patent is a breakthrough using the Kelly et al., 2021 data. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee. The sample is a subset of the full sample, including all control teams, and teams treated by the death of a team member who did replace the deceased team member.

FIGURE B.7
TREATMENT EFFECT ACROSS TEAM SPAN VOLUME



Notes: This figure plots the marginal effect of the treatment variable on the probability a team's patent is a breakthrough. The coefficients are taken from the regression outlined in B.8.

TABLE B.8
BREAKTHROUGH RESULTS: INTERACTION MODEL

	Dependent variable: Pr(Breakthrough)
$D_{\tau t}$	0.0047**
	(3.06)
Volume$_{\tau}$	-2.7874***
	(-3.97)
$D_{\tau'} \times$ Volume$_{\tau}$	-0.0028**
	(-3.03)
Prior work$_{\tau_1 t}$	-0.0131***
	(-4.86)
N	9325
Controls	✓
Team FE	✓
Period FE	✓
Year FE	✓

Notes: The model used is a conditional logit with team and patent order fixed effects, where standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable is an indicator for whether the patent is a breakthrough using the Kelly et al., 2021 data. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee. The sample is a subset of the full sample, including all control teams, and teams treated by the death of a team member who did replace the deceased team member.

This paper designs the treatment model around the premature death of inventors. This provides exogenous variation in team composition, and therefore the team's position in knowledge space. To demonstrate the robustness of the results in the paper I include a set of teams which add a new inventor. This gives a weakly larger knowledge field for the team, and potentially allows them to build on more or different types of prior work.

I confirm the robustness of these results by finding that teams which add a new member, and increase the number of patents targeting a specific direction see a significant increase in the probability they patent in that direction. The reverse holds for the breakthrough patents, though the results are weaker.

TABLE B.9
TEAM TREATMENT ESTIMATES: KELLY ET AL., 2021

Dependent variable: Pr(Breakthrough)				
	1.	2.	3.	4.
$D_{\tau t}$	0.0009***	0.0041***	0.0032**	0.0027*
	(3.39)	(4.07)	(2.89)	(2.38)
Prior work$_{\tau_1 t}$	-0.0015***	-0.0114***	-0.0076***	-0.0075***
	(-10.68)	(-6.91)	(-5.98)	(-5.91)
Volume$_{\tau}$				-0.6563
				(-1.04)
N	25812	5863	5863	5863
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable is an indicator for whether the patent is a breakthrough using the Kelly et al., 2021 data. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

TABLE B.10
TEAM TREATMENT ESTIMATES: ADDING AN INVENTOR
I: BREAKTHROUGH

Dependent variable: Pr(Breakthrough)				
	1.	2.	3.	4.
$D_{\tau t}$	-0.0029**	-0.0048	-0.0008	-0.0027
	(-2.71)	(-1.46)	(-0.74)	(-1.11)
Prior work$_{\tau_1 t}$	-0.0008***	-0.0264***	-0.0178***	-0.0175***
	(-10.18)	(-5.36)	(-4.60)	(-4.55)
Volume$_{\tau}$				0.9827
				(1.53)
N	29349	9560	9560	9560
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period FE		✓	✓	✓
Year FE			✓	✓

II: DIRECTION

Dependent variable: Pr(Direction)				
	1.	2.	3.	4.
$D_{\tau t}$ | Direction	0.0130***	0.0093***	0.0016	0.0023
	(6.56)	(3.79)	(0.81)	(1.06)
Prior work$_{\tau_1 t}$ | Direction	0.0088***	0.0614***	0.0294***	0.0293***
	(18.87)	(11.16)	(12.07)	(12.07)
Volume$_{\tau}$				-0.2062
				(-1.05)
N	88047	61755	61755	61755
Controls	✓	✓	✓	✓
Team FE		✓	✓	✓
Period × Direction FE		✓	✓	✓
Year × Direction FE			✓	✓

Notes: The first column presents a standard logit model. Columns 2-4 are conditional logit models with team and patent order fixed effect models and standard errors are clustered at this level. The identifier τ is unique for each team pair (τ_1, τ_2). The dependent variable for panel I) is an indicator for whether the patent is a breakthrough. The breakthrough classification is based on equation 2.7, which defines it as the post-count (patents produced in the field after the given patent) normalized by the sum of the post-count and prior count (patents in the field before the given patent). The dependent variable for panel II) is a stacked indicator for whether a patent achieves the given direction: mitigates climate change, reduces cancer risk or automates production. Controls include $d(\theta_p^e, \theta_p)$, team gender diversity, average team experience and its squared term, race diversity and the rolling three year average number of inventors employed at the patent assignee.

C. Chapter 3 Appendix

C.1 Theory

Proof. To prove proposition 1, we proceed in two steps. First, we prove the proposition for a much simpler setting with with two firms and one local labor market in which they overlap. Second, we show how the result in step one is easily established in the general model.

Consider the maximization problem of firm j, which has a degree of Common Ownership with firm $-j$ with which also shares a local labor market,

$$\begin{aligned}
&\max \quad p_j Y_j - rK_j - w(L)L_j + \lambda \left(p_{-j} Y_{-j} - rK_{-j} - w\left(L_j + L_{-j}\right) L_{-j}\right) \\
&\text{subject to} \\
&Y_j = \exp\left(\int_0^1 \ln\left[y_j(x)\right] dx\right)^{\nu} \\
&y_j(x) = \gamma_m(x) m_j(x) + \gamma_l(x) l_j(x) \\
&K_j = \int_0^I m_j(x) dx \\
&L_j = \int_I^1 l_j(x) dx
\end{aligned} \tag{C.1}$$

Given the assumption regarding the comparative advantage structure, the FOCs of the problem are,

$$\begin{aligned}
&[m(x)] \Longrightarrow m(x) = \frac{Y}{r}\nu \\
&[\ell(x)] \Longrightarrow \ell(x) = \frac{Y}{W}\nu \\
&[I] \Longrightarrow \frac{\gamma_\ell\left(I_j\right)}{\gamma_m\left(I_j\right)} = \frac{W}{r}
\end{aligned} \tag{C.2}$$

with W being equal to the marginal cost of labor, e.g.,

$$W = w(L) + w'(L)(L_j + \lambda L_{-j}). \tag{C.3}$$

Rearranging the FOCs, we obtain that the solution to the problem of the firms is the solution to the two-equation two-unknowns problem. The system of equa-

tions is,[1]

$$\begin{aligned} L_j &= (1 - I_j) \left(\frac{W}{\nu}\right)^{\frac{I_j \nu - 1}{1-\nu}} \left[G \left(\frac{\nu}{r}\right)^{I_j}\right]^{\frac{\nu}{1-\nu}} \\ W &= w(L) + w'(L)(L_j + \lambda L_{-j}). \end{aligned} \tag{C.4}$$

and the unknowns are L_j and W. Recall that both I_j and G are also functions of W. We need to prove that as Common Ownership, λ, increases, employment decreases, and automation increases. We will focus on proving that employment decreases and the marginal cost of labor W increases. Indeed, an increase in W is necessarily linked with an increase in automation I_j—see the FOC in C.2. To proceed with the analytical proof, we need to specify the functional forms of the productivity schedules. These are,

$$\gamma_m(x) = e^{\alpha_m x} \tag{C.5}$$

$$\gamma_\ell(x) = e^{\alpha_\ell x} \tag{C.6}$$

with, crucially, $\alpha_\ell > \alpha_m$. This implies that the expression for I_j is,

$$I = \frac{1}{\alpha_\ell - \alpha_m} \log\left(\frac{W}{r}\right) \tag{C.7}$$

To characterize the solution of the system in C.4, we first prove that the function represented by the first equation, $L_j = g_d(W)$ is decreasing, that is, as the marginal cost of labor goes up, the labor demand decreases. To begin, take logs,

$$\begin{aligned} \log(L) = \log(1 - I) + \frac{I\nu}{1-\nu}(\log(W) - \log(\nu) + \log(\frac{\nu}{r}))- \\ \frac{1}{1-\nu}\log(W) + \frac{1}{1-\nu}\log(\nu) + \frac{\nu}{1-\nu}\log(G) \end{aligned} \tag{C.8}$$

take the derivative with respect to W

$$\frac{\partial}{\partial W}\log(L) = \frac{\partial I}{\partial W}\left[\frac{\nu}{1-\nu}\log(\frac{W}{r}) - \frac{1}{1-I}\right] + \frac{I\nu - 1}{(1-\nu)W} + \frac{\nu}{1-\nu}\frac{1}{G}\frac{\partial G}{\partial W} \tag{C.9}$$

now substitute the derivative of G

$$\frac{\partial G}{\partial W} = -G \log\left(\frac{W}{r}\right)\frac{\partial I}{\partial W} \tag{C.10}$$

[1] Recall that

$$G = \exp\left(\int_0^I \log[\gamma_m(x)]dx + \int_I^1 \log[\gamma_\ell(x)]dx\right)$$

and obtain,

$$\begin{aligned}\frac{\partial}{\partial W}\log(L) = \frac{\partial I}{\partial W}\left[\frac{\nu}{1-\nu}\log(\frac{W}{r}) - \frac{1}{1-I}\right] \\ + \frac{I\nu - 1}{(1-\nu)W} - \frac{\nu}{1-\nu}\log(\frac{W}{r})\frac{\partial I}{\partial W}\end{aligned} \tag{C.11}$$

by rearranging we obtain,

$$\frac{\partial}{\partial W}\log(L) = -\frac{\partial I}{\partial W}\frac{1}{1-I} + \frac{I\nu - 1}{(1-\nu)W} \tag{C.12}$$

Which is always negative because $0 < I < 1$ and $0 < \nu < 1$. As the function $L_j = g_d(W)$ is decreasing, the inverse $W = g_d^{-1}(L_j)$ exists and is also decreasing. As the second equation of the system C.4 is increasing, we can plot the two functions and derive the properties of the solution. In Figure C.1 we plot two scenarios. One in which the level of Common Ownership is λ^1 and a second one with higher Common Ownership, $\lambda^2 > \lambda^1$. The change in CO does not affect the labor demand *curve*. As CO shifts the labor supply curve upwards, and because of the proven properties of the labor demand curve, the optimal employment of firm j decreases, and the marginal cost of labor goes up. As can be easily seen by looking at equation (C.7), an increase in W_j implies an increase in I_j.

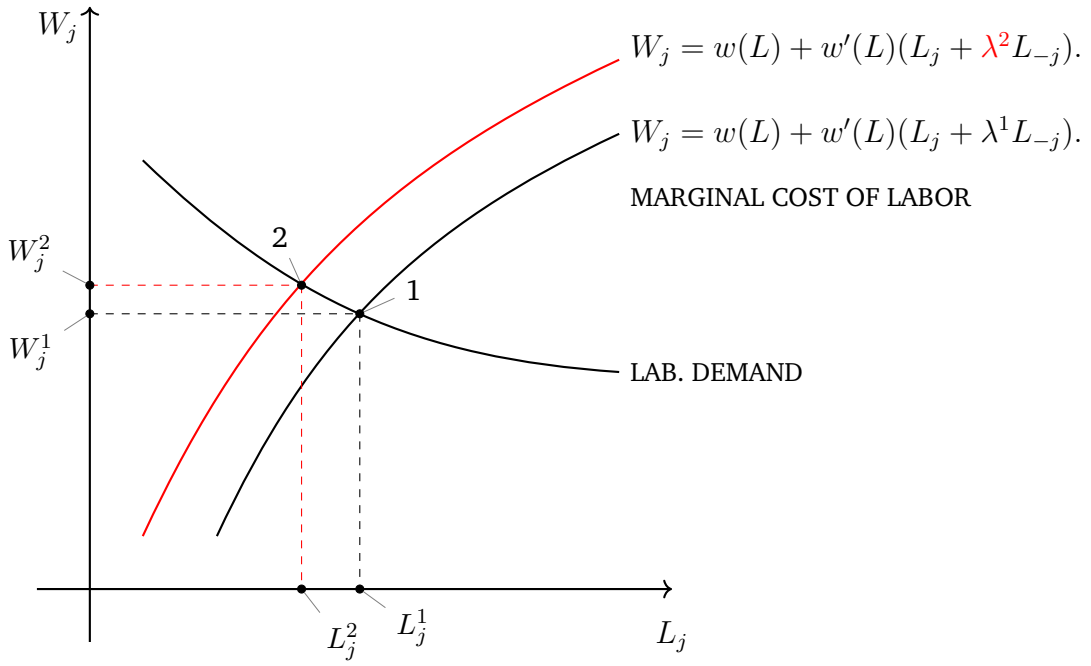
It is straightforward to generalize this result to the model with a generic number of firms and local labor markets. In the general model, an increase in a pairwise degree of Common Ownership λ_{fj} increases the level of automation in the local labor market I_f^c which consequently, increases the automation *average* across local labor markets,

$$I_f \equiv \frac{1}{|\mathbf{C_f}|}\sum_{\mathbf{C_f}} I_f^c. \tag{C.13}$$

We can provide a visual proof for proposition 1 through the following figure. This provides a more intuitive grasp on the proof discussed. This figure plots the labor demand curve and the curve of the marginal factor cost of labor for two different values of common ownership, $\lambda_2 > \lambda_1$. It shows that as common ownership goes up employment decreases and the marginal cost of labor increases.

TABLE C.1

A VISUAL PROOF FOR PROPOSITION 1



Notes: This figure plots the labor demand curve and the curve of the marginal factor cost of labor for two different values of common ownership where $\lambda_2 > \lambda_1$.

□

C.2 OLS Estimation

TABLE C.1

OLS REGRESSION RESULTS

	(1) Y_{t+1}	(2) Y_{t+2}	(3) Y_{t+3}	(4) Y_{t+4}	(5) Y_{t+5}	(6) Y_{t+6}
C Index	9.851	12.610*	16.750**	17.741**	14.898	8.571
	(6.513)	(7.029)	(7.637)	(8.613)	(9.496)	(10.532)
InstOwn	0.026	-0.026	-0.007	0.061	0.045	-0.041
	(0.073)	(0.074)	(0.076)	(0.077)	(0.082)	(0.085)
R&D	0.049	0.043*	0.013	0.001	-0.019	0.025
	(0.038)	(0.023)	(0.025)	(0.021)	(0.035)	(0.030)
Firm Size	0.064***	0.052***	0.040**	0.027	0.018	0.029
	(0.021)	(0.020)	(0.019)	(0.018)	(0.020)	(0.020)
Firm Age	-0.108**	-0.082*	-0.108**	-0.088*	-0.123**	-0.160***
	(0.043)	(0.044)	(0.045)	(0.046)	(0.048)	(0.048)
PPE / Assets	-0.123	-0.052	0.027	-0.032	0.050	0.043
	(0.105)	(0.111)	(0.114)	(0.118)	(0.120)	(0.126)
Constant	-0.033	-0.024	0.083	0.091	0.233	0.320**
	(0.151)	(0.145)	(0.143)	(0.142)	(0.149)	(0.151)
Observations	19725	18163	16562	14989	13477	12033
R^2_{adj}	0.742	0.749	0.758	0.769	0.775	0.785

*Notes: This Table presents OLS estimates of firms' automation strategy on common ownership with LLM rivals. Standard errors in parenthesis are clustered at the firm level. Significance levels are indicated as follows: * p<0.1, ** p<0.05, *** p<0.01.*

C.3 Robustness

C.3.1 Using Citation-Weighted Patents

In studies concerning innovation, it is standard to control for the quality of patents by using citation counts. We also have estimated our models concerning innovation outcomes on the intensive margin using citation-weighted patents, applying the "time-technology class fixed effect" method (Hall, Jaffe, and Trajtenberg, 2001; Atanassov, 2013), to address truncation problems.

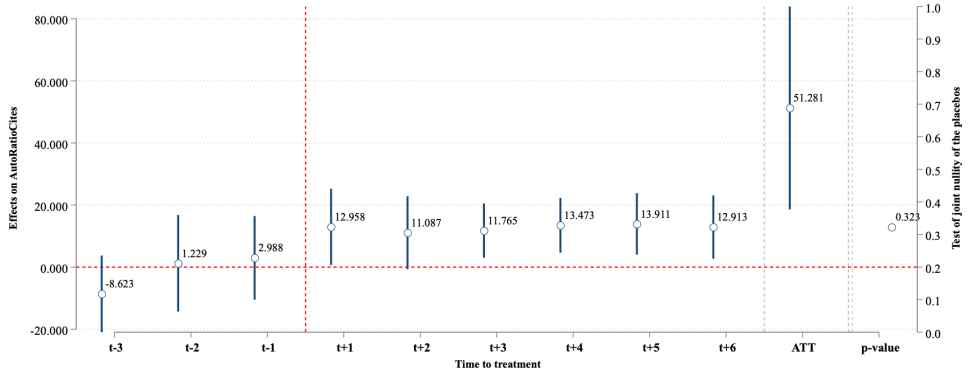
With these adjusted patent citations, we compute our continuous innovation outcome variables. That is,

$$\textit{AutoRatioCites} = \ln\left[\frac{1 + \text{citation-weighted automation Patents}}{1 + \text{citation-weighted non-automation Patents}}\right],$$

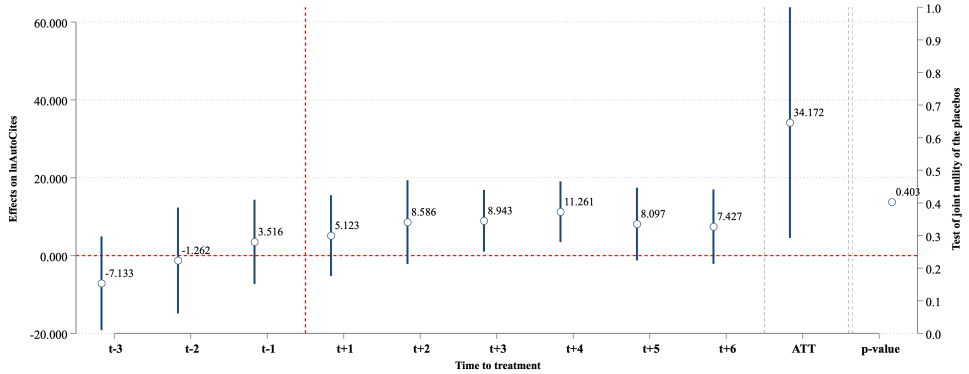
as well as *lnAutoCites* and *lnNonAutoCites*, which are the natural logarithm of (one plus) the citation-weighted number of Patents for automation and non-automation, respectively. The results are shown in Figure C.2.

TABLE C.2

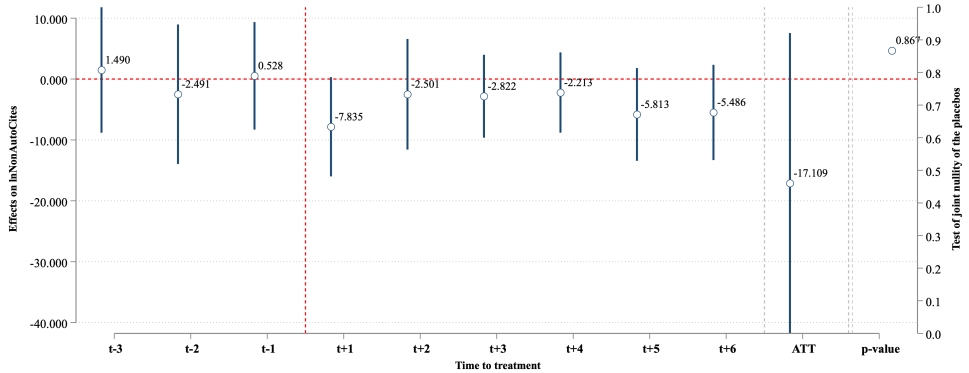
DYNAMIC EFFECT: COMMON OWNERSHIP WITH WEIGHTED PATENT COUNTS



(a) Using *AutoRatioCites* as dependent variable.



(b) Using *lnAutoCites* as dependent variable.



(c) Using *lnNonAutoCites* as dependent variable.

Notes: This Figure shows the dynamic effects of within-LLM increases to common ownership (Cont-TreatLLM) on the automation strategy of firms based on citation-weighted patent counts.

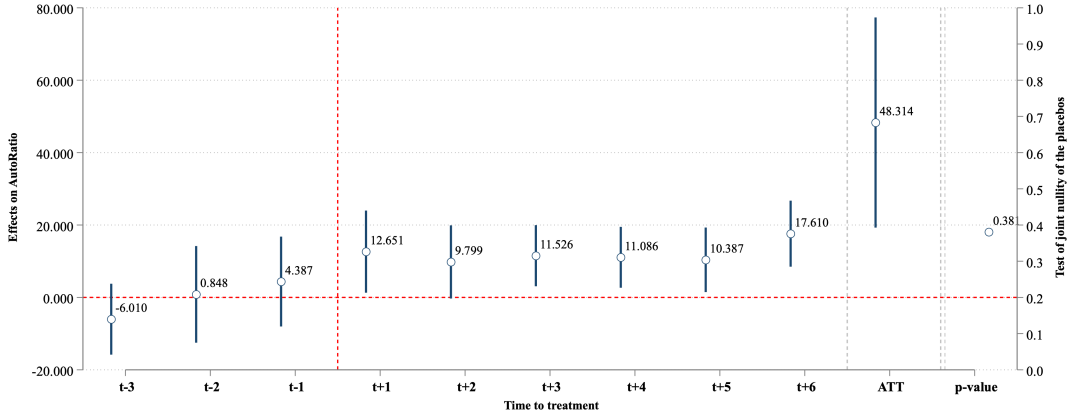
C.3.2 Using Data Pre-Financial Crisis

In this section, we address the concern that our results may be driven by abnormalities during the financial crisis. We present the same analysis as in the

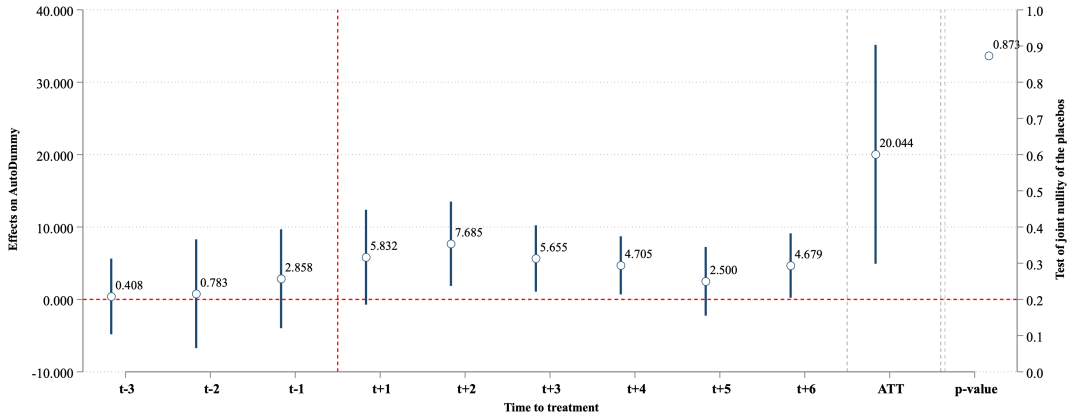
main text of our paper, but using only data up to 2006. That means, we also exclude the last seven mergers in the sample of institutional mergers identified by Lewellen and Lowry, 2021, including the merger between BlackRock and Barclays Global Investors, which was used for identification in previous studies.[2] The results are shown in Figure C.3.

TABLE C.3

DYNAMIC EFFECTS: AUTOMATION UNTIL 2006



(a) Using the continuous measure *AutoRatio*



(b) Using the binary measure *AutoDummy*

Notes: This Figure shows the dynamic effects of within-LLM increases to common ownership (ContTreatLLM) on the automation strategy of firms using data until 2006, i.e., before the onset of the financial crisis.

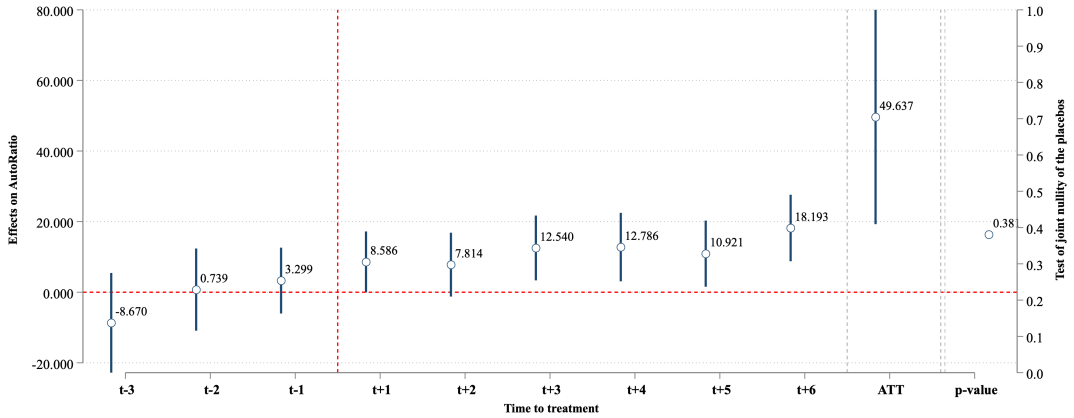
C.3.3 Binary Treatment Setup

We present the results of the DID analysis employing a binary treatment variable *TreatLLM* as defined in Section 3.3.2. The results are shown in Figure C.4.

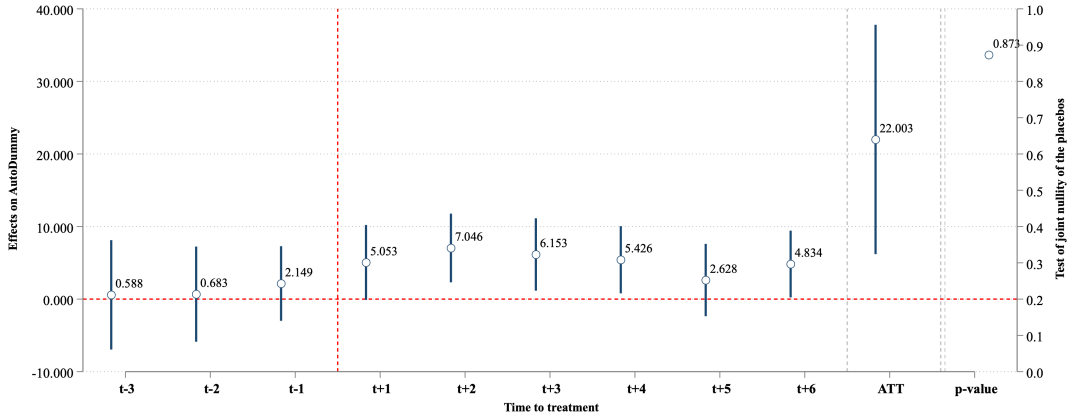
[2] See, e.g., Azar, Schmalz, and Tecu, 2018.

TABLE C.4

DYNAMIC EFFECTS: AUTOMATION WITH DISCRETE TREATMENT



(a) Using the continuous measure *AutoRatio*



(b) Using the binary measure *AutoDummy*

Notes: This Figure shows the dynamic effects of within-LLM increases to common ownership using a binary treatment variable (TreatLLM) on the automation strategy of firms.

C.4 Database Construction

Matching establishment level information with Compustat. The 2020 version of the National Establishments Time Series (NETS) provides the legal business names of establishments. We employ the legal business names of these establishments for cross-referencing with the Compustat database in our empirical analysis.

We utilize fuzzy matching, employing a similarity threshold of 90%, to align company names between the two databases. Subsequently, we conduct manual verification to ensure the precision of these matches. With this methodology, we successfully merge 353,818 establishments. It results in a dataset of 4,231,721 establishment-year observations for the spanning period 1990-2020, each containing no missing information on employee count and geographical location.

Bibliography

Abowd, **John M.**, **Francis Kramarz**, and **David N. Margolis** (1999). "High Wage Workers and High Wage Firms". *Econometrica* 67.2, pp. 251–333.

Acemoglu, **Daron**, **Claire Lelarge**, and **Pascual Restrepo** (2020). "Competing with robots: Firm-level evidence from France". *AEA papers and proceedings*. Vol. 110, pp. 383–388.

Acemoglu, **Daron**, **Andrea Manera**, and **Pascual Restrepo** (2020). *Does the US tax code favor automation?* Tech. rep. National Bureau of Economic Research.

Acemoglu, **Daron** and **Pascual Restrepo** (2018). "The race between man and machine: Implications of technology for growth, factor shares, and employment". *American Economic Review* 108.6, pp. 1488–1542.

– (2020). "Robots and jobs: Evidence from US labor markets". *Journal of Political Economy* 128.6, pp. 2188–2244.

Aghion, **Philippe**, **Céline Antonin**, **Simon Bunel**, and **Xavier Jaravel** (2020). "What are the labor and product market effects of automation? New evidence from France". *CEPR Discussion Paper*.

Aghion, **Philippe**, **Antoine Dechezleprêtre**, **David Hémous**, **Ralf Martin**, and **John Van Reenen** (2016). "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry". *Journal of Political Economy* 124.1, pp. 1–51.

Aghion, **Philippe**, **John Van Reenen**, and **Luigi Zingales** (2013). "Innovation and institutional ownership". *American economic review* 103.1, pp. 277–304.

Agrawal, **Ajay**, **Avi Goldfarb**, and **Florenta Teodoridis** (Jan. 2016). "Understanding the Changing Structure of Scientific Inquiry". *American Economic Journal: Applied Economics* 8.1, pp. 100–128.

Ahmadpoor, **Mohammad** and **Benjamin F. Jones** (2019). "Decoding team and individual impact in science and invention". *PNAS* 116.28, pp. 13885–13890.

Akcigit, **Ufuk**, **Santiago Caicedo**, **Ernest Miguelez**, **Stefanie Stantcheva**, and **Valerio Sterzi** (2018). *Dancing with the Stars: Innovation Through Interactions*. Working Paper 24466. National Bureau of Economic Research.

Akerman, **Anders**, **Ingvil Gaarder**, and **Magne Mogstad** (2015). "The skill complementarity of broadband internet". *The Quarterly Journal of Economics* 130.4, pp. 1781–1824.

Anderson, **Katharine A.** (2012). "Specialists and generalists: Equilibrium skill acquisition decisions in problem-solving populations". *Journal of Economic Behavior and Organization* 84.1, pp. 463–473.

Anton, **Miguel**, **Florian Ederer**, **Mireia Gine**, and **Martin C Schmalz** (2018). "Innovation: the bright side of common ownership?" *Available at SSRN 3099578*.

Arntz, **Melanie**, **Terry Gregory**, and **Ulrich Zierahn** (2016). "The risk of automation for jobs in OECD countries: A comparative analysis".

Arora, **Ashish**, **Sharon Belenzon**, and **Lia Sheer** (Mar. 2021). "Knowledge Spillovers and Corporate Investment in Scientific Research". *American Economic Review* 111.3, pp. 871–98.

Arts, **Sam**, **Jianan Hou**, and **Juan Carlos Gomez** (2021). "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures". *Research Policy* 50.2, p. 104144.

Arts, **Sam**, **Nicola Melluso**, and **Reinhilde Veugelers** (Jan. 2025). "Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text". *The Review of Economics and Statistics*, pp. 1–33.

Atanassov, **Julian** (2013). "Do Hostile Takeovers Stifle Innovation? Evidence from Antitakeover Legislation and Corporate Patenting". *Journal of Finance* 68.3, pp. 1097–1131.

Autor, **David H** and **David Dorn** (2013). "The growth of low-skill service jobs and the polarization of the US labor market". *American economic review* 103.5, pp. 1553–1597.

Autor, **David H**, **Frank Levy**, and **Richard J Murnane** (2003). "The skill content of recent technological change: An empirical exploration". *The Quarterly Journal of Economics* 118.4, pp. 1279–1333.

Azar, **José**, **Marina Chugunova**, **Klaus Keller**, and **Sampsa Samila** (2023). "Monopsony and Automation". *Max Planck Institute for Innovation & Competition Research Paper* 23-21.

Azar, **José**, **Yue Qiu**, and **Aaron Sojourner** (2022). "Common ownership in labor markets". *Available at SSRN*.

Azar, **José**, **Sahil Raina**, and **Martin Schmalz** (2022). "Ultimate ownership and bank competition". *Financial Management* 51.1, pp. 227–269.

Azar, **José**, **Martin C Schmalz**, and **Isabel Tecu** (2018). "Anticompetitive effects of common ownership". *The Journal of Finance* 73.4, pp. 1513–1565.

Azar, **José** and **Xavier Vives** (2019). "Common ownership and the secular stagnation hypothesis". *AEA Papers and Proceedings*. Vol. 109. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 322–326.

– (2021). "General equilibrium oligopoly and ownership structure". *Econometrica* 89.3, pp. 999–1048.

Azoulay, **Pierre**, **Christian Fons-Rosen**, and **Joshua S. Graff Zivin** (2019). "Does Science Advance One Funeral at a Time?" *American Economic Review* 109.8, pp. 2889–2920.

Azoulay, **Pierre**, **Joshua S. Graff Zivin**, and **Jialan Wang** (May 2010). "Superstar Extinction". *The Quarterly Journal of Economics* 125.2, pp. 549–589.

Backus, **Matthew**, **Christopher Conlon**, and **Michael Sinkinson** (2021). "Common ownership in America: 1980–2017". *American Economic Journal: Microeconomics* 13.3, pp. 273–308.

Baker, **Jonathan B** (2015). "Overlapping financial investor ownership, market power, and antitrust enforcement: My qualified agreement with Professor Elhauge". *Harv. L. Rev. F.* 129, p. 212.

Bandiera, **Oriana**, **Andrea Prat**, **Stephen Hansen**, and **Raffaella Sadun** (2020). "CEO Behavior and Firm Performance". *Journal of Political Economy* 128.4, pp. 1325–1369.

Battaglia, **Laura**, **Timothy Christensen**, **Stephen Hansen**, and **Szymon Sacher** (2024). "Inference for Regression with Variables Generated from Unstructured Data". *Unpublished manuscript*.

Beaudry, **Paul**, **Mark Doms**, and **Ethan Lewis** (2010). "Should the personal computer be considered a technological revolution? Evidence from US metropolitan areas". *Journal of political Economy* 118.5, pp. 988–1036.

Ben-David, **Itzhak**, **Francesco Franzoni**, **Rabih Moussawi**, and **John Sedunov** (2016). *The granular nature of large institutional investors*. Tech. rep. National Bureau of Economic Research.

Bessen, **James**, **Martin Goos**, **Anna Salomons**, and **Wiljan Van den Berge** (2023). "Automatic reaction-what happens to workers at firms that automate?" *The Review of Economics and Statistics* Feb. 6, 2023.

Blei, **David M** and **John D Lafferty** (2005). "Correlated topic models". *NeurIPS*, pp. 147–154.

Blei, **David M.**, **Andrew Y. Ng**, and **Michael I. Jordan** (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3, pp. 993–1022.

Bloom, **Nicholas**, **Charles I. Jones**, **John Van Reenen**, and **Michael Webb** (2020). "Are Ideas Getting Harder to Find?" *American Economic Review* 110.4, pp. 1104–44.

Bloom, **Nicholas**, **Mark Schankerman**, and **John Van Reenen** (2013). "Identifying technology spillovers and product market rivalry". *Econometrica* 81.4, pp. 1347–1393.

Boerma, **Job**, **Aleh Tsyvinski**, and **Alexander P Zimin** (2021). *Sorting with Team Formation*. Working Paper 29290. National Bureau of Economic Research.

Bonfiglioli, **Alessandra**, **Rosario Crino**, **Harald Fadinger**, and **Gino Gancia** (2024). "Robot imports and firm-level outcomes". *The Economic Journal*.

Bonhomme, **Stéphane** (2022). “Teams: Heterogeneity, Sorting, and Complementarity”. *Unpublished manuscript*.

Brusoni, **Stefano** and **Andrea Prencipe** (Mar. 2001). “Unpacking the Black Box of Modularity: Technologies, Products and Organizations”. *Industrial and Corporate Change* 10.1, pp. 179–205.

Cancer Moonshot: USPTO (2024). *Cancer Moonshot Patent Data*.

Carvalho, **Vasco M.**, **Mirko Draca**, and **Nikolas Kuhlen** (Aug. 2021). “Exploration and Exploitation in US Technological Change”.

Danzer, **Alexander M**, **Carsten Feuerbaum**, and **Fabian Gaessler** (2024). “Labor supply and automation innovation: Evidence from an allocation policy”. *Journal of Public Economics* 235, p. 105136.

Dauth, **Wolfgang**, **Sebastian Findeisen**, **Jens Suedekum**, and **Nicole Woessner** (2019). “The adjustment of labor markets to robots”. *Journal of the European Economic Association*.

De Chaisemartin, **Clément** and **Xavier d’Haultfoeuille** (2024). “Difference-in-differences estimators of intertemporal treatment effects”. *Review of Economics and Statistics*, pp. 1–45.

Deng, **Liuchun**, **Steffen Müller**, **Verena Plümpe**, and **Jens Stegmaier** (2024). “Robots, occupations, and worker age: A production-unit analysis of employment”. *European Economic Review (forthcoming)*.

Devereux, **Kevin** (2018). *Identifying the value of teamwork: Application to professional tennis*. Working Paper Series 14. University of Waterloo.

Ethiraj, **Sendil K.** and **Daniel Levinthal** (2004). “Modularity and Innovation in Complex Systems”. *Management Science* 50.2, pp. 159–173.

Fleming, **Lee** (2001). “Recombinant Uncertainty in Technological Search”. *Management Science* 47.1, pp. 117–132.

Fleming, **Lee** and **Olav Sorenson** (2004). “Science as a map in technological search”. *Strategic Management Journal* 25.8-9, pp. 909–928.

Freund, **Lukas** (2022). *Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities*. Working Paper. SSRN.

Frey, **Carl Benedikt** and **Michael A Osborne** (2017). "The future of employment: How susceptible are jobs to computerisation?" *Technological forecasting and social change* 114, pp. 254–280.

Friedman, **Milton** and **Anna J. Schwartz** (1963). *A Monetary History of the United States, 1867-1960*. Princeton, NJ: Princeton University Press.

Gingras, **Yves**, **Vincent Larivière**, **Benoît Macaluso**, and **Jean-Pierre Robitaille** (2008). "The effects of aging on researchers' publication and citation patterns". *PLoS ONE* 3.12.

Goos, **Maarten** and **Alan Manning** (2007). "Lousy and lovely jobs: The rising polarization of work in Britain". *The Review of Economics and Statistics* 89.1, pp. 118–133.

Graetz, **Georg** and **Guy Michaels** (2018). "Robots at work". *Review of Economics and Statistics* 100.5, pp. 753–768.

Graves, **Jennifer** and **Zoë Kuehn** (2021). "Specializing in growing sectors: Wage returns and gender differences". *Labour Economics* 70, p. 101994.

Griffiths, **Thomas L.** and **Mark Steyvers** (2004). "Finding Scientific Topics". *Proceedings of the National Academy of Sciences* 101.1, pp. 5228–5235.

Grossman, **Gene M** and **Elhanan Helpman** (1991). "Quality Ladders in the Theory of Growth". *Review of Economic Studies* 58.1, pp. 43–61.

Guo, **Bing**, **Dennis C. Hutschenreiter**, **David Pérez-Castrillo**, and **Anna Toldrà-Simats** (2024). "Institutional Blockholders and Corporate Innovation".

Guo, **Bing**, **David Pérez-Castrillo**, and **Anna Toldrà-Simats** (2019). "Firms' innovation strategy under the shadow of analyst coverage". *Journal of Financial Economics* 131.2, pp. 456–483.

Haeussler, **Carolin** and **Henry Sauermann** (2020). "Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity". *Research Policy* 49.6, p. 103987.

Hall, **Bronwyn**, **Manuel Trajtenberg**, and **Adam B. Jaffe** (2001). *The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools*. Working Paper 3094. Centre for Economic Policy Research.

Hall, **Bronwyn H**, **Adam B Jaffe**, and **Manuel Trajtenberg** (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*.

Hansen, **Stephen**, **Michael McMahon**, and **Andrea Prat** (2018). "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach". *Quarterly Journal of Economics* 133.2, pp. 801–870.

He, **Jie** and **Jiekun Huang** (2017a). "Product market competition in a world of cross-ownership: Evidence from institutional blockholdings". *The Review of Financial Studies* 30.8, pp. 2674–2718.

He, **Jie (Jack)** and **Jiekun Huang** (Apr. 2017b). "Product Market Competition in a World of Cross-Ownership: Evidence from Institutional Blockholdings". *The Review of Financial Studies* 30.8, pp. 2674–2718. ISSN: 0893-9454.

Herkenhoff, **Kyle**, **Jeremy Lise**, **Guido Menzio**, and **Gordon M. Phillips** (2024). "Production and Learning in Teams". *Econometrica* 92.2, pp. 467–504.

Hutschenreiter, **Dennis** (2023). "Common Ownership and the Market for Technology".

Hutschenreiter, **Dennis C**, **Tommaso Santini**, and **Eugenia Vella** (2022). "Automation and sectoral reallocation". *SERIEs* 13.1, pp. 335–362.

Hutschenreiter, **Dennis C.** and **Tommaso Santini** (2021). "Common Ownership and Automation". *Santini, Three essays on automation. PhD Thesis*.

Jaffe, **Adam B** (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value". *American Economic Review* 76.5, pp. 984–1001.

Jayaraman, **Narayanan**, **Ajay Khorana**, and **Edward Nelling** (2002). "An analysis of the determinants and shareholder wealth effects of mutual fund mergers". *The Journal of Finance* 57.3, pp. 1521–1551.

Jones, **Benjamin** (2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *The Review of Economic Studies* 6 (1), pp. 283–317.

Kahane, **Leo**, **Neil Longley**, and **Robert Simmons** (2013). "The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and

Language Diversity in the National Hockey League". *The Review of Economics and Statistics* 95.1, pp. 302–314.

Kaltenberg, **Mary**, **Adam B Jaffe**, and **Margie Lachman** (2021). *The Age of Invention: Matching Inventor Ages to Patents Based on Web-scraped Sources*. Working Paper 28768. National Bureau of Economic Research.

Kaltenberg, **Mary**, **Adam B. Jaffe**, and **Margie E. Lachman** (2023). "Invention and the life course: Age differences in patenting". *Research Policy* 52.1, p. 104629.

Kelly, **Bryan**, **Dimitris Papanikolaou**, **Amit Seru**, and **Matt Taddy** (2021). "Measuring Technological Innovation over the Long Run". *American Economic Review: Insights* 3.3, pp. 303–20.

Koch, **Michael**, **Ilya Manuylov**, and **Marcel Smolka** (2021). "Robots and firms". *The Economic Journal* 131.638, pp. 2553–2584.

Kogan, **Leonid**, **Dimitris Papanikolaou**, **Amit Seru**, and **Noah Stoffman** (Mar. 2017). "Technological Innovation, Resource Allocation, and Growth*". *The Quarterly Journal of Economics* 132.2, pp. 665–712.

Lewellen, **Katharina** and **Michelle Lowry** (2021). "Does common ownership really increase firm coordination?" *Journal of Financial Economics* 141.1, pp. 322–344. ISSN: 0304-405X.

López, **Ángel L** and **Xavier Vives** (2019). "Overlapping ownership, R&D spillovers, and antitrust policy". *Journal of Political Economy* 127.5, pp. 2394–2437.

Macho-Stadler, **Ines** and **Thierry Verdier** (1991). "Strategic managerial incentives and cross ownership structure: a note". *Journal of Economics* 53.3, pp. 285–297.

Mann, **Katja** and **Lukas Püttmann** (Aug. 2021). "Benign Effects of Automation: New Evidence from Patent Texts". *The Review of Economics and Statistics*, pp. 1–45. ISSN: 0034-6535.

– (May 2023). "Benign Effects of Automation: New Evidence from Patent Texts". *The Review of Economics and Statistics* 105.3, pp. 562–579.

Melero, **Eduardo** and **Neus Palomeras** (2015). "The Renaissance Man is not dead! The role of generalists in teams of inventors". *Research Policy* 44.1, pp. 154–167.

Michaels, **Guy**, **Ashwini Natraj**, and **John Van Reenen** (2014). "Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years". *Review of Economics and Statistics* 96.1, pp. 60–77.

Mindruta, **Denisa**, **Janet Bercovitz**, **Vlad Mares**, and **Maryann Feldman** (2024). "Stars in Their Constellations: Great Person or Great Team?" *Management Science*.

Moll, **Benjamin**, **Lukasz Rachel**, and **Pascual Restrepo** (2022). "Uneven growth: automation's impact on income and wealth inequality". *Econometrica* 90.6, pp. 2645–2683.

Mortensen, **Olavur** (2017). "The Author Topic Model". *Unpublished manuscript*.

PatentsView (2024). *USPTO Patent Data for Inventors and Assignees*. United States Patent and Trademark Office (USPTO).

Pearce, **Jeremy** (2022). "Idea Production and Team Structure". *Unpublished manuscript*.

Posner, **Eric A**, **Fiona M Scott Morgan**, and **E Glen Weyl** (2016). "A proposal to limit the anticompetitive power of institutional investors". *Antitrust LJ* 81, p. 669.

Rosen-Zvi, **Michal**, **Thomas L. Griffiths**, **Mark Steyvers**, and **Padhraic Smyth** (2012). "The Author-Topic Model for Authors and Documents". *CoRR* abs/1207.4169, pp. 487–494.

Santini, **Tommaso** (2024). *Automation with heterogeneous agents: the effect on consumption inequality*. Tech. rep. IWH Discussion Papers.

Sarica, **Serhad** and **Jianxi Luo** (2020). "Stopwords in Technical Language Processing". *CoRR* abs/2006.02633.

Sauermann, **Henry** and **Carolin Haeussler** (2017). "Authorship and contribution disclosures". *Science Advances* 3.11.

Schumpeter, **Joseph A.** (1939). *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*. New York: McGraw-Hill.

Singh, **Jasjit** and **Lee Fleming** (2010). "Lone Inventors as Sources of Breakthroughs: Myth or Reality?" *Management Science* 56.1, pp. 41–56.

Sun, **Liyang** and **Sarah Abraham** (2020). *Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects*.

Teodoridis, **Florenta**, **Jino Lu**, and **Jeffrey L Furman** (2022). *Mapping the Knowledge Space: Exploiting Unassisted Machine Learning Tools*. Working Paper 30603. National Bureau of Economic Research.

Terry, **Stephen J**, **Thomas Chaney**, **Konrad B Burchardi**, **Lisa Tarquinio**, and **Tarek A Hassan** (2024). "Immigration, Innovation, and Growth".

Uzzi, **Brian**, **Satyam Mukherjee**, **Michael Stringer**, and **Ben Jones** (2013). "Atypical Combinations and Scientific Impact". *Science* 342, pp. 468–472.

Vakili, **Keyvan** and **Sarah Kaplan** (2021). "Organizing for innovation: A contingency view on innovative team configuration". *Strategic Management Journal* 42.6, pp. 1159–1183.

von Hippel, **Eric** (1990). "Task partitioning: An innovation process variable". *Research Policy* 19.5, pp. 407–418.

Weidmann, **Ben** and **David J. Deming** (2021). "Team Players: How Social Skills Improve Team Performance". *Econometrica* 89.6, pp. 2637–2657.

Weitzman, **Martin L.** (1998). "Recombinant Growth". *The Quarterly Journal of Economics* 113.2, pp. 331–360.

Wu, **Lingfei**, **Dashun Wang**, and **James A Evans** (2019). "Large teams develop and small teams disrupt science and technologyt". *Nature* 566, pp. 378–382.

Wu, **Renli**, **Christopher Esposito**, and **James Evans** (2024). *China's Rising Leadership in Global Science*. Working Paper 2406.05917. ArXiv.

Wuchty, **Stefan**, **Benjamin F. Jones**, and **Brian Uzzi** (2007). "The Increasing Dominance of Teams in Production of Knowledge". *Science* 316.5827, pp. 1036–1039.

Xu, **Fengli**, **Lingfei Wu**, and **James A Evans** (2013). “Flat teams drive scientific innovation”. *PNAS* 119.23.