Teams and Text: Cover Note

This cover note accompanies my submission of the paper *Teams and Text: Disentangling Team Knowledge Production* for the Innovation Information Initiative Technical Working Group Meeting, Fall 2025.

The paper introduces a new model to estimate individual inventors' contribution shares within team projects. The framework disentangles each team member's input and estimates a contribution share for each team member. This allows for a more precise mapping of how knowledge is organised within teams, offering a valuable tool for research on science and innovation. I have previously applied the method in my working paper *Collaborative Innovation in the Knowledge Space*. Indeed, the methodology was originally developed as part of my Job Market Paper, where it appeared as a technical appendix.

I have included a basic replication package on my GitHub for the past two years and recently developed a significantly improved implementation. The workshop would be an ideal setting to refine the paper and gather feedback on making the repository a useful resource for the community. My plan is currently to provide a single data run alongside the replication package on publication, enabling parameter tuning and adaptation to alternative data. Though this is, of course, where feedback discussions could be very productive.

I believe the methodology opens up a wide range of research questions on inventor compensation, team diversity, the knowledge production function and the rewards to learning, among many others.

P.S. I wanted to highlight, that unfortunately I wouldn't be unable to attend on December 5th, due to a clash with a conference I am jointly organising at MIT. I of course understand if preference is given to participants who can attend both days.

Thank you for your consideration.

Joseph Emmens

Teams and Text: Disentangling Team Knowledge Production

Joseph Emmens[‡]

MIT FutureTech

September 23, 2025

Click here for the latest version.

Abstract

This paper develops a novel Bayesian natural language processing framework to recover the division of labour in inventor teams directly from patent texts. The method estimates the probability that each inventor contributed specific knowledge components, informed by their prior collaborations. Applying this framework to approximately 4 million patents, I validate the approach and introduce a novel measure of individual contributions within teams. The results show that as inventors gain experience, they collaborate on more patents per year. However, within projects, senior colleagues contribute relatively less than their junior collaborators. Importantly, concentrating contributions in the hands of lower-performing team members significantly reduces patent value, highlighting how the structure of collaboration shapes the social returns to innovation.

Keywords: Teams, Collaboration, Innovation, Patents, Text Analysis

JEL: O31 O32 O34 C55

[‡]Massachusetts Institute of Technology—FutureTech. An earlier version of this paper was circulated as part of the appendix to the paper "Teams and Text: Collaborative Innovation in the Knowledge Space". I am grateful to Sam Arts, Matt Clancy, Christian Fons-Rosen, Fabian Gaessler, Jacob Hess, Manuel Lleonart Anguix, Luis Ignacio Menéndez Garcia, Hannes Mueller, Neus Palomeras, Olav Sorensen, Weiliang Tan, Florenta Teodoridis and Sergi Quintana for their helpful discussion and comments. Part of this project was developed through funding provided by the Spanish Ministry of Science and Innovation Grant: PID2020-114251GB-I00. Website: www.josephemmens.com. Email: jemmens@mit.edu

1 Introduction

Teams drive scientific and technological progress, yet how they organise the production of knowledge remains largely a black box. A perennial concern in team production is the contribution of individual team members. When a breakthrough technology is developed, who gets the most recognition within a team? Furthermore, individual recognition is essential for incentives in team environments. Consider Dr. Falana, who developed a novel method for making chiral additive chemicals. His employer later used this technology to make the specific compounds essential to LCD screens with U.S. Patent 6,830,789, "Chiral Additives for Cholesteric Displays".[1] Though Dr. Falana was initially omitted from the list of inventors, a court reviewing the patent documents ruled that his methodological contribution to the innovation was essential, and added him as co-inventor. Thus awarding Dr. Falana the corresponding financial and reputational gains associated with being named as an inventor.

This paper presents a novel method of disentangling contributions to the knowledge production process and shows that this division can be inferred from patent texts. The following definition of contribution is used: the share of knowledge components contained in an innovation contributed by a given inventor. For example, if a self-driving car was developed using knowledge from computer vision, machine learning and automotive engineering we might be curious which team member provided expertise in which area. Did one inventor work alone on the perception algorithms, while another focused on vehicle mechanics, or did all members collaborate equally to all parts?[2]

To infer these contribution shares, I model collaboration through the lens of the Author-Topic Model (Rosen-Zvi et al., 2012). This method models the probability that a team member contributed each word, representing a specific knowledge component, where this probability is informed by their patent history with other teams. If an inventor with a history of conventional car innovations appears on a patent for a self-driving car, alongside co-inventors with histories in machine learning, the model allocates them a larger weight for the vehicle mechanics, than the perception algorithms.

To illustrate, I apply this model to a dataset of nearly 4 million USPTO patents. I find that as inventors gain experience, they collaborate on more patents per year. Within each project, however, I show that senior inventors tend to contribute less than their junior collaborators.

[1] Olusegun Falana v. Kent State University, 669 F.3d 1349 (Federal Circuit, 2012).

[2] It is important to note that it does not reflect the effort contributed to other facets of collaboration: data curation, applying for grants, and other administrative tasks; In this sense, the paper is different from Xu, Wu, and Evans (2013), who use a definition of contribution to general scientific work load.

This matters for patent value: patents where contributions are concentrated on one or a few team members receive significantly fewer citations. This result is consistent with evidence from literature that flat teams drive science (Xu, Wu, and Evans, 2013). At an aggregate level, teams divide contributions differently across technological areas, and I demonstrate that teamwork is getting "flatter" over time.

The contribution of this paper lies in developing a novel model that allows for the scientific division of labour within inventor teams. The framework provides new insights into the structure of innovative collaboration, while presenting a powerful yet parsimonious framework to model the knowledge production function. To validate the model, I show that when contribution shares are highly unequal, the lead inventor's CPC technological class history provides up to 25% more information for predicting the class awarded to the team's patent. Given this evidence of having captured an important economic variable, the framework can be applied to questions linking contribution shares with wages, occupations, and recognising the contribution of women and minorities in science and technology. Therefore opening a broad agenda for future research in the economics of science and innovation.

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; Jones, 2023). 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 facilitates the inference of 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 an 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 contributions. 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. I lean on a similar topic model to Teodoridis, Lu, and Furman (2022) and Carvalho, Draca, and Kuhlen (2021) however moving from the patent to team level to understand the key producers of innovation. This therefore places the paper in the wider literature using patent and paper texts to develop novel empirical measures of scientific and technological progress (Kelly et al., 2021; Arts, Hou, and Gomez, 2021; Marx and Fuegi, 2022; Arts, Melluso, and Veugelers, 2025).

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.

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.[3] 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 and provides a validation test; Section 5 presents results and section 6 concludes.

[3]They show that this result holds for normalising by team size, which, given trends of increasing team sizes, this normalisation may introduce an alternative explanation.

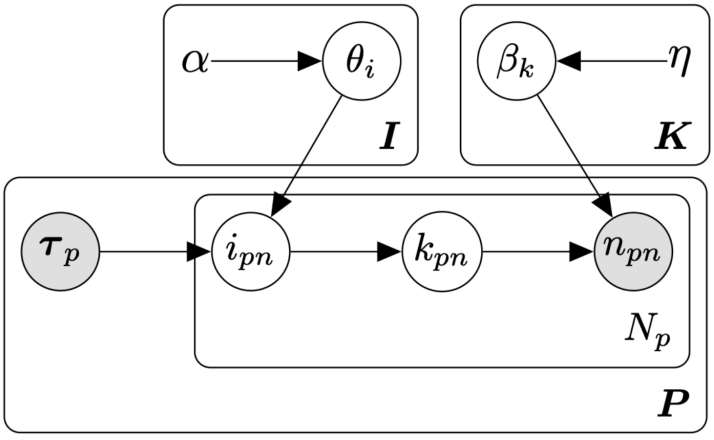
2 Disentangling Teamwork Through Text

In order to disentangle individual contributions to team projects, I present the Author-Topic Model as a novel model of collaboration. This model is a hierarchical Bayesian model of Natural Language Processing. The first version of this 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.[4]

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

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. α and η are the hyper-parameters chosen by the researcher. Variables shaded in gray are observed, including the team τ and patent words n_{pn}. The model proposed maps teams to text through the inventors i and knowledge classes k where each are drawn from the inventor to knowledge class distributions θ_i and knowledge class to word distributions β_k.

[4]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.

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. The model assumes a uniform prior contribution share for each team member of $1/m$. This is updated to a non-uniform posterior after observing the data. This model is represented in the following plate diagram in Figure (1).

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

$$P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \alpha, \eta, \boldsymbol{n}, \boldsymbol{T}) = \frac{P(\boldsymbol{n}|\boldsymbol{k}, \boldsymbol{\beta}) P(\boldsymbol{k}|\boldsymbol{i}, \boldsymbol{\Theta}) P(\boldsymbol{i}|\boldsymbol{T}) P(\boldsymbol{\beta}|\eta) P(\boldsymbol{\Theta}|\alpha)}{P(\boldsymbol{n}|\alpha, \eta, \boldsymbol{T})} \tag{1}$$

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.[5] Define $q(\cdot)$ as an approximation to the posterior

$$q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \lambda, \gamma, \phi) = q(\boldsymbol{\Theta}|\boldsymbol{\gamma}) q(\boldsymbol{\beta}|\lambda) q(\boldsymbol{k}, \boldsymbol{i}|\phi) \tag{2}$$

$$\approx P(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \alpha, \eta, \boldsymbol{n}, \boldsymbol{T}) \tag{3}$$

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

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 2 now models the choice of knowledge classes and inventors as dependent random variables 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 $q(\cdot)$

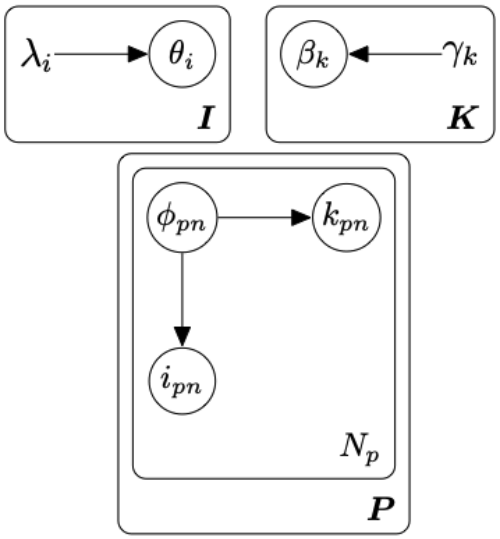
$$\begin{aligned} q(\boldsymbol{k}, \boldsymbol{i}, \boldsymbol{\beta}, \boldsymbol{\Theta}|\lambda, \gamma, \phi) &= q(\boldsymbol{\Theta}|\boldsymbol{\gamma})q(\boldsymbol{\beta}|\lambda)q(\boldsymbol{k}, \boldsymbol{i}|\phi) \\ &= \prod_i q(\theta_i|\gamma_i) \prod_k q(\beta_k|\lambda_k) \prod_{p,n}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

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 (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

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

Figure 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. The variational parameters are λ which governs the inventor to knowledge class distribution and γ the knowledge class to word distribution. ϕ now governs the joint distribution over both inventors i and knowledge classes k.

The full probability distribution is stored during the estimation as a four dimensional matrix $\phi_{p,v,i,k}$[6]. Where $\sum_i \sum_k \phi_{p,v,i,k} = 1$.

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.[7]

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

[6]In reality the Mortensen, 2017 package uses the exchangeability of the model to develop an online algorithm to reduce the memory requirements of this matrix

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

and optimising in that direction, holding all others constant until convergence.

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

On convergence, I back out the θ_i given γ_i and β_k given λ_k.[8] 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

$$\phi_{p,v,i,k} = \frac{\exp\{\mathbb{E}_q[\log\,\theta_{ik}] + \mathbb{E}_q[\log\,\beta_{kv}]\}}{\sum_k \sum_{i\in\boldsymbol{\tau}_p} \exp\{\mathbb{E}_q[\log\,\theta_{ik}] + \mathbb{E}_q[\log\,\beta_{kv}]\}} \tag{4}$$

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 $v \in V$.

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}

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

will increase. This can be seen through equation 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.

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 $\bar{\omega}_m = \frac{1}{m}1m$, 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.

$$\text{Concentration}_{\tau p} = ||\bar{\omega}_m - \omega_p||^2 \tag{5}$$

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.

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 take the universe of USPTO patents which were applied for, and later granted, between the years 1976 and 2024. I clean out the continuation patents, using the USPTO parent to child patent mapping. To ensure a long patent history for each inventor, and to keep computational requirements manageable, I slice the sample in the following way. First, I take the giant component on the inventor network. I extend this by only taking patents produced by inventors who patent more than five patents over their career. 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. This leads to a sample of 3751439 patents written by 788799 inventors to train the model of Latent Dirichlet Allocation. Many studies utilising patent texts use the abstract. As a concise description, it serves the purpose of describing the innovation and the idea behind it well. They are, however, structured as a list of components, where I require the distribution across these components in order to learn the contribution share of each inventor. Therefore, I require the full patent text. In order to keep the computational cost down, I take a random selection of 1000 words from the patent text, drawing each with uniform probability.

For the regression analysis in section 5 I drop all teams of one inventor. These teams are very useful for measuring an inventor's knowledge profile, however, do not provide meaningful variation on which to describe how team's distribute contribution. The final sample is described in Table 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 converting all text to lower-case and filtering out punctuation. Each patent text is then tokenised to split it into individual words. Each token

Table 1
DESCRIPTIVE STATISTICS

Variaable	Observations	Mean	Std. Dev.	Min	Max
		Patent Level			
Citations	2,192,414	13.537	48.699	0	3582
Year	2,192,414	2005.948	10.726	1976	2024
Concentration	2,192,414	0.085	0.078	0.000	0.711
Patent Experience	2,192,414	22.601	17.995	2	373
		Inventor Level			
Number of Patents	657,933	13.495	27.135	1	2143
Number of Teams	657,933	7.392	12.868	1	1040
Inventor Quality	641,260	0.015	0.210	−2.341	3.457
Years Experience	3,262,327	6.234	5.178	1	35
Yearly Patents	3,262,327	2.884	4.374	1	1330
Normalised Contribution Share	6,931,882	1.000	0.204	0.006	6.512
Average Co-inventor Quality	6,931,882	7.148	4.575	1	35
		Team Level			
Team Size	1,350,570	3.213	1.526	2	38
Number of Patents	1,350,570	1.623	1.829	1	327

Notes: The table reports the mean, standard deviation, minimum, and maximum values. This is the sample used in the main analysis, not to train the LDA. This sample drops all teams of size 1. Observations represent the number of patents, inventors, or teams in each category. Inventor quality is measured as as the fixed effect from a Abowd, Kramarz, and Margolis, 1999 regression of their contribution share multiplied by the citation count at the patent level on a set of inventor and team dummies. All variables are presented in the raw count and not the log transformation used, for some, in the empirical analysis.

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 stop words, which carry little semantic value due to their regular use in patents and scientific texts, alongside words shorter than three characters.[9]

Table 2 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 10000. The choice of $\eta = 1/K = 0.025$ is in line with Griffiths and Steyvers (2004) and Hansen, McMahon, and Prat (2018). I estimate the Bayesian parameter flexibly, rather

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

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 technology 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.

Table 2

LDA Parameters

K	η	Iterations	Passes	γ
77	0.025	50	20	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 parameters during the expectation maximisation step. 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.*

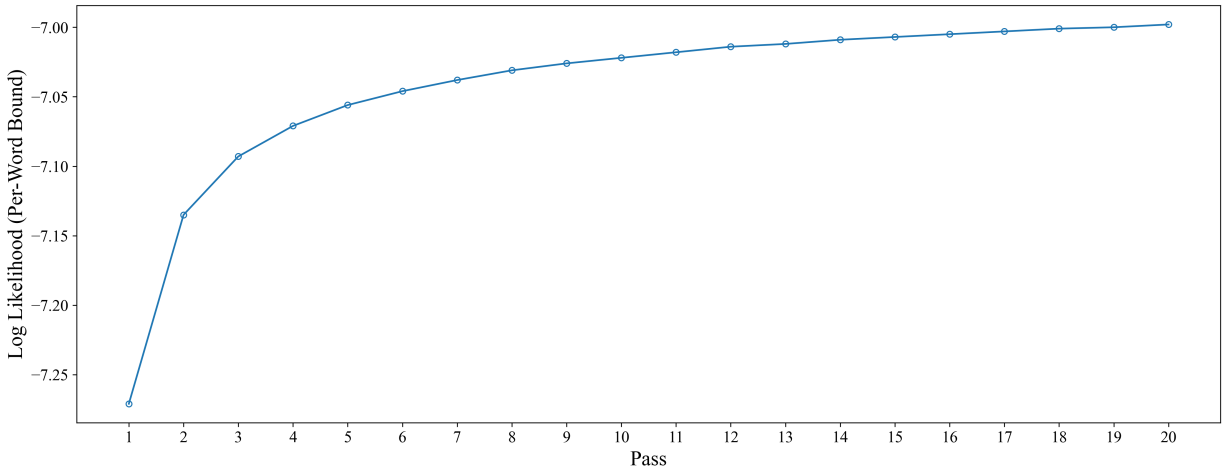
4.1 Convergence

The log-likelihood per-word bound is a common statistic used to measure convergence in topic models. 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. Figure 3 shows that the ATM converges after around 20 passes.

4.2 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

Figure 3
LDA Model Convergence



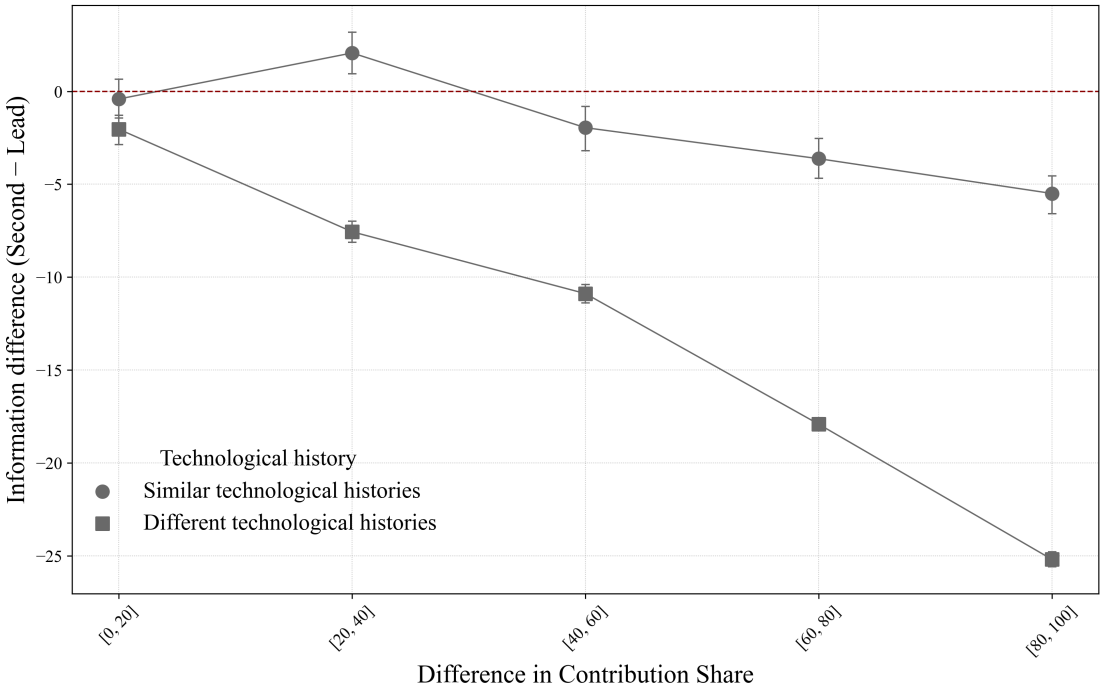
Notes: A plot showing the model convergence in the log-likelihood as a per-word bound. The per-word log likelihood is the average log likelihood of each word under the model, showing how well the model explains the data. Normalising by the number of words is standard in the Topic Model literature.

contribution weights using a prediction model. I propose that if the weights capture information on the true contribution share of each 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.

For each patent in the sample, I define the lead and last inventors 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 ten most common CPC classes used by the lead inventor prior to the target patent, and the corresponding ten for the last inventor. Ten is the maximum number taken, if either of the inventors have fewer than ten, then I take the minimum between both inventors to ensure an equal number of features for both. The feature importance presented is the average value, therefore uneven technological histories does not drive these results. 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

Figure 4
CONTRIBUTION SHARE VALIDATION



Notes: A plot of the mean difference between the feature importance for the lead and last inventor on a given patent, when predicting a patent's CPC classification. The lead and last inventor are the team member with the largest and smallest contribution share, respectively. The prediction model is a random forest. For each bin I bootstrap the random forest training and test split, allowing me to average out the random nature of a random forest predictor and generate the confidence intervals. I run 300 bootstrap runs, with 100 trees in each random forest, each with a maximum depth of 20.

patenting history being significantly larger than that of the second inventor. I extend this by splitting the patents into two groups, those in which the first and last inventor's have different technological histories, and those for which they have similar. I do this by calculating the Jaccard similarity between the technological histories of each inventor. The distribution of these scores are given in Figure B2, different technological histories are those with a score below 0.5. Which in other words means that less than half the technology classes that belong to the union of histories belong to the intersection.

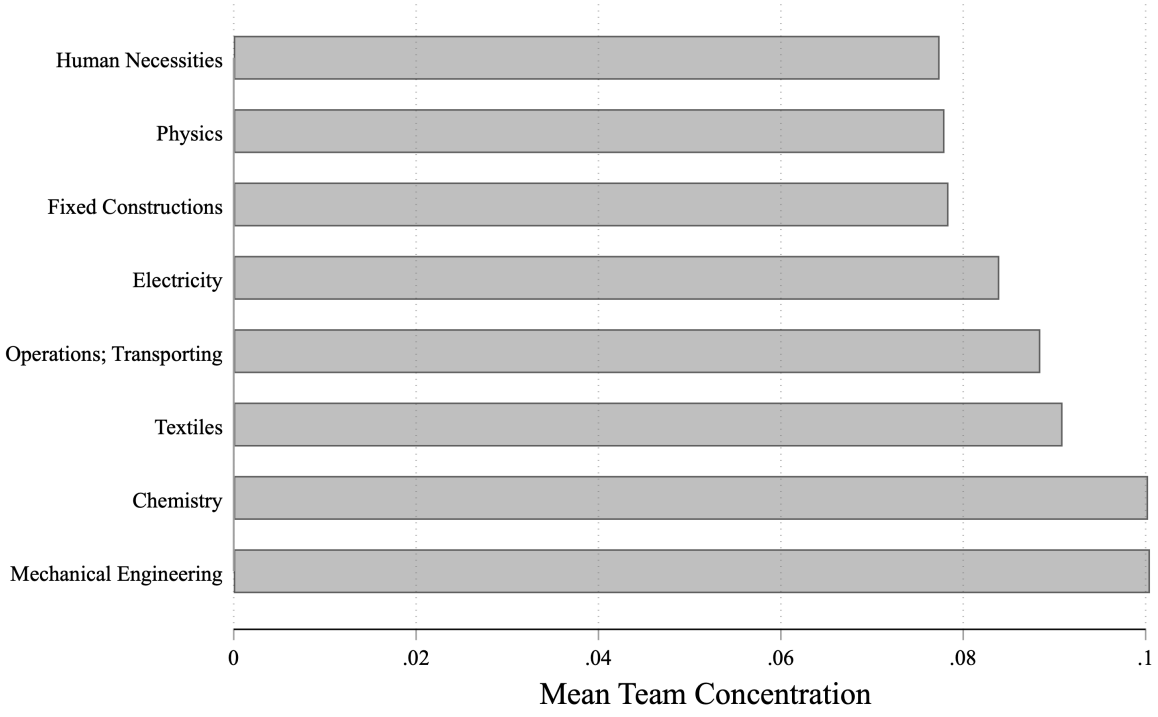
Figure 4 plots the difference in feature importance between the first and last inventor's patenting histories, across binned groups of the difference between their inferred ATM-contribution share, split by both similar and different technological histories. The feature difference is defined as the feature 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. I show that on patents for which the lead and last inventor have similar histories, when the contribution gap is below 60%, the difference in the information provided by either inventor's technological history hovers around zero. Only when the gap between their contributions grows very large does a gain of up to 5 percentage points more information appear for the lead inventors technological history.

There is however a strong negative trend for teams in which the technological histories for both inventors are different. The greater the difference in the inferred ATM contribution share, the more information the leading inventor's patenting history provides in the prediction task. For the largest difference in contribution shares the lead inventor provides up to 25% more information than the last. This points to the contribution weights providing economically important and precise information on who contributed to the knowledge contained.

Figure 5

Concentration over Technology Area



Notes: 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 within each team. Bars are sorted by mean concentration, with labels corresponding to CPC sections. CPC titles are abridged. N = 1,916,541.

5 Results

I begin by presenting a set of novel descriptive statistics that examine concentration across technology classes and time. I present a set of aggregate descriptive statistics for the concentration measure introduced in equation 5.

Figure 5 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 6 I fit a linear polynomial using the Epanechnikov kernel and calculate the confidence interval at the 95% level. The average concentration across the patents produced in each year. This result shows that on aggregate, teams are getting flatter.

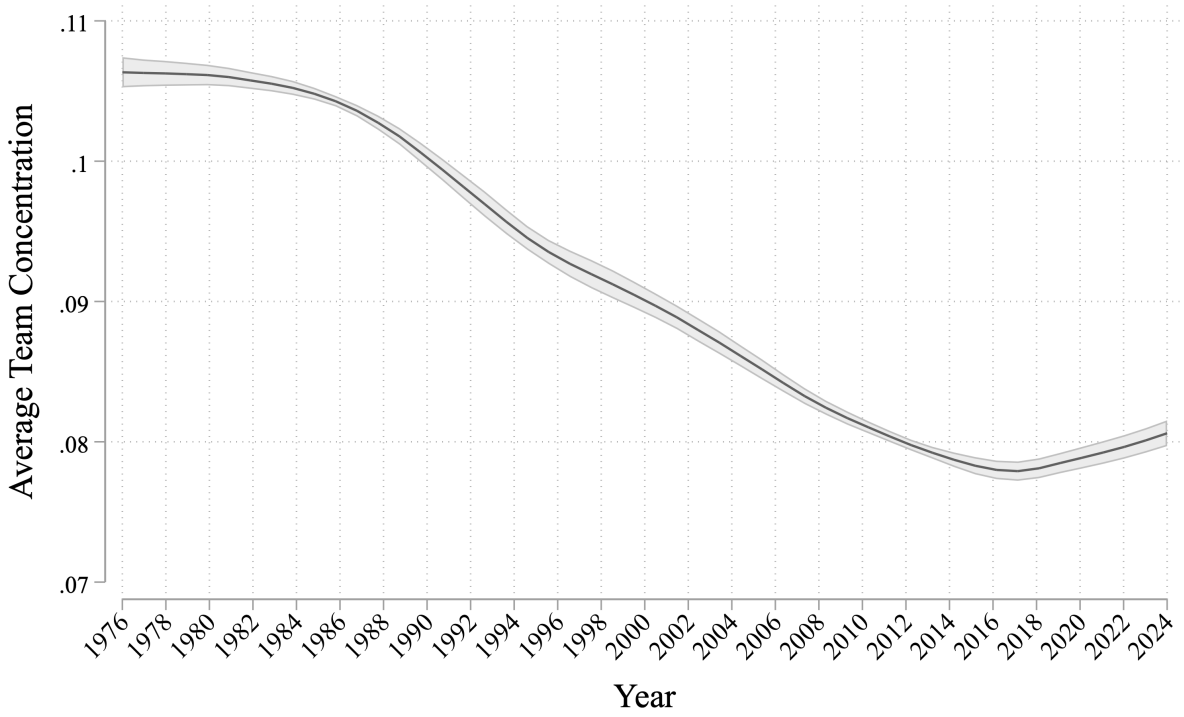
5.1 Results on Contribution Shares

By analysing individual contributions to the knowledge components of a patent, I uncover new insights into how teams structure their knowledge production. First of all, I show that the number of patents an inventor collaborates on each year increases with each additional year of patenting experience. Y_{it} is the number of patents inventor i appears on in year t. Formally, the regression is given by:

$$Y_{it} = \alpha_i + \sum_{s=1}^{35} \beta_s \cdot \mathbb{1}(\text{years experience}_{it} = s) + \delta_t + \epsilon_{it} \tag{6}$$

The regression includes a set of dummy variables capturing whether inventor i in year t has exactly s years of patenting experience, as well as a set of year dummies. The result is shown in Figure B4, which plots the predicted number of patents collaborated on per year of patenting experience, after cleaning out inventor fixed effects. Given that this sample is of inventors with at east five patents over their career, this sample over represents prolific inventors compared to the universe.

Figure 6
CONCENTRATION OVER TIME



Notes: This figure presents the mean concentration of inventor teams from 1976-2024, measured by the Euclidean distance between the vector of contribution shares and a uniform distribution within each team. The x-axis labels plot every other year for conciseness. I fit a linear polynomial using the Epanechnikov kernel and calculate the confidence interval at the 95% level. N = 1,916,541.

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, I define the outcome variable as a contribution share normalised by an equal share for a team of size m.

$$Y_{ip} = \frac{\omega_{ip}}{1/m}$$

This aids in the interpretation of any predicted values. If an inventor's normalised contribution share is below one, they are not contributing their "fair share", vice versa for those above one. I examine this question through two regression models, each presenting a complementary view of how knowledge production is organised within teams.

In the first specification, I use a patent fixed effect to isolate how variations in team member characteristics shape the way contributions are organised. I find that more senior inventors

tend to contribute a lower share than their more junior co-inventors. This is shown through the following regression and Figure B5.

$$Y_{ip} = \alpha_p + \sum_{s=1}^{35} \beta_s \cdot \mathbb{1}(\text{years experience}_{it} = s) + \epsilon_{ip} \tag{7}$$

The patent fixed effect holds the average experience of the team constant, therefore by increasing that of the focal inventor they are becoming relatively more senior. Those who are have accumulated 35 years of patenting experience, on average, contribute over 30% less than an equal share baseline.

This result emerges most starkly when inventor i's relative experience is compared directly with that of their co-inventors. To do so, I split the variable tracking years patenting into three equally sized bins to track the experience of inventor i on patent p ($exp_{i(p)}$). These correspond to junior, middle and senior experience levels. Juniors have fewer than 4 years experience, Middle have 5-8 and Seniors have between 8-35 years of experience.

I build a second count, binning the average number of years of experience patenting for the inventor's collaborators in team τ, prior to the patent p. I denote the team τ minus inventor i by $\tilde{\tau}$. I split this same variable into the same three bins ($exp_{\tilde{\tau}(p)}$). This allows me to track whether inventor i collaborated with junior or senior co-inventors. I include indicator functions for inventor and average co-inventor experience levels, and their interaction. I again control for patent fixed effects which hold a large range of potential patent value confounding variables constant.

$$\begin{aligned} Y_{ip} = \alpha_p &+ \sum_{e=1}^{3} \beta_e \cdot \mathbb{1}(\exp_{i(p)} = e) + \sum_{o=1}^{3} \beta_o \cdot \mathbb{1}(\exp_{\tilde{\tau}(p)} = o) \\ &+ \sum_{e=1}^{3} \sum_{o=1}^{3} \delta_{eo} \cdot \mathbb{1}(\exp_{i(p)} = e) \times \mathbb{1}(\exp_{\tilde{\tau}(p)} = o) + \epsilon_{ip} \end{aligned}$$

Figure B6 replicates the previous result that senior inventors tend to contribute less when working with junior colleagues. This effect is strongest when a senior collaborates with junior inventors. An inventor tends to contribute closer to their equal share when they hold the same experience level as their average co-inventor.

To directly compare an inventors contribution share to the number of patents they collaborate on each year, I re-run the model for a within-inventor model. An important caveat here is

that selection will drive part of these results, since the inventors who accumulate many years of patenting experience are likely to not be a representative sample. While an inventor fixed effect will control for time invariant traits, there may be dynamic effects, such as learning management practices, driving these results. Despite this, these results are of value given that we can examine the set of inventors that return to patent in significantly more detail than previously.

Given that the dependent variable now varies at the inventor × year level, I aggregate patents for an inventor within each year as the average normalised contribution share. The regression is specified as

$$\bar{Y}_{it} = \alpha_i + \sum_{s=1}^{35} \beta_s \cdot \mathbb{1}(\text{years experience}_{it} = s) + \mu \cdot \text{co-inventor experience}_{it} + \delta_t + \epsilon_{ip} \quad (8)$$

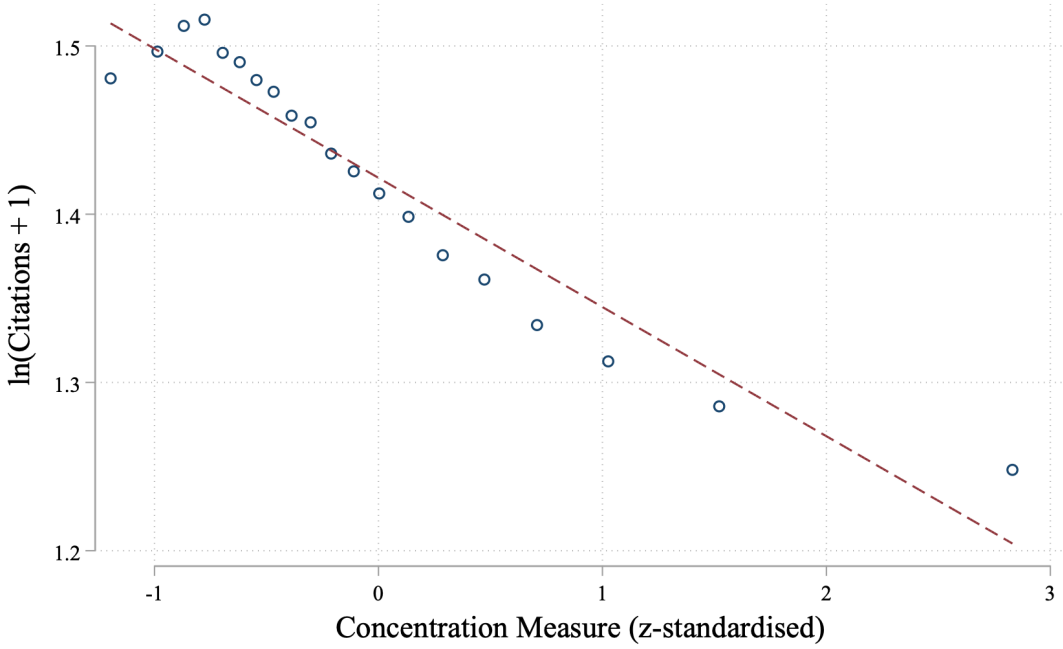
The results are conditional on the experience of the team with which the inventor is collaborating, through the variable co-inventor experience$_{it}$. Figure B8 shows that, on average, inventors contribute to patents as they gain an additional year of experience. This would contradict a narrative of them working on more patents each year, but contributing less to each, though should be interpreted in the context of the selection previously described. I directly test this hypothesis through Figure B7, in which I now predict an inventors contribution share over the number of patents they are collaborating on in a given year. For a clear answer at the within-inventor level, these results need to be updated to handle the selection problem described.

5.2 Results on Patent Value

Regardless of which factor drives inventors to contribute relatively more or less within a team, I demonstrate that the scientific division of labour within teams matters for patent value. If an individual's contribution decreases, this must be compensated by increasing the contribution of one or a few other team members. This translates into a higher concentration score as defined in equation 5. Figure 7 shows that concentration has a strong negative correlation with patent value, measured through the natural logarithm of citations received. The slope of the fitted line is given in Table 3, in column 1. There is potentially an inverted-U shape, with some gains at very low levels of concentration. However, the shape from the

mean level onwards is clear, concentration is negatively associated with patent value.

Figure 7
PATENT VALUE AND CONTRIBUTION CONCENTRATION



Notes: A bin-scatter plot of patent value, measured as the natural logarithm of the number of citations a patent receives plus one, on the concentration of contribution shares within the team. Concentration is measured by the euclidean distance between the vector of observed contribution weights for a team on a given patent, minus the vector of equal contribution weights for a team of that size. The concentration score is z-standardised. I control for a set of technology class, year and team size dummy variables as well as controlling for the level of experience of the team.

To test this relationship further, I control for the contribution share allocated to the top inventor in the team. Where the top inventor is the inventor with the highest quality in the team. Inventor quality is measured through regression, in the style of Abowd, Kramarz, and Margolis (1999), using both inventor (α_i) and team (α_τ) fixed effects.

$$\tilde{y}_{ip} = \alpha_i + \alpha_\tau + \delta_t + \epsilon_{ip} \tag{9}$$

where $\tilde{y}_{ip}$ is the natural logarithm of number of citations received by patent p weighted by the contribution share of inventor i to patent p, plus one. This coefficient identifies the average increase to citations received from increasing the contribution weight of the given inventor. Therefore this measure of quality is static, in that each team has one unique ordering of team member quality.

The identification in this model is therefore across teams, as opposed to within teams.

Perhaps unsurprisingly, when contribution is loaded on high quality inventors, patent value increases, as can be seen through the coefficient on Top Inventor Share in Table 3.

To examine this directly, I introduce the interaction between the quality of the top inventor and their contribution share. Clearly, teams with better top inventors produce patents of higher value. However, as shown through this model, this effect is only realised when you load contribution share onto that inventor. Across all regression models I control for how experienced the team is, and find a consistent negative coefficient. This is inline with the argument that senior, and potentially older, inventor's produce less impactful innovations.

Table 3

Regression Model Results: Patent Value

	(1)	(2)	(3)	(4)
	Dependent variable: *ln(Citations+1)*			
Concentration	-0.090***	-0.129***	-0.147***	-0.153***
	(0.001)	(0.001)	(0.001)	(0.001)
Team Experience	-0.011***	-0.012***	-0.011***	-0.011***
	(0.000)	(0.000)	(0.000)	(0.000)
Top Inventor Share		0.394***	0.378***	0.373***
		(0.005)	(0.006)	(0.006)
Top Inventor Quality			0.112***	-0.000
			(0.001)	(0.006)
(Top Inventor) Share × Quality				0.099***
				(0.005)
Technology Class × Year Fixed Effects	✓	✓	✓	✓
R^2	0.481	0.483	0.487	0.488
N	1,916,541	1,916,541	1,872,972	1,872,972

Notes: For all regressions the dependent variable is the natural logarithm of citations received plus one. Both the concentration and lead inventor quality measures are z-standardised. All four regressions control for a set of technology class × year dummy variables. The variable (Top Inventor) Share × Quality corresponds to the interaction between the variable Top Inventor Share and Top Inventor Quality. Standard errors are clustered at the team level. t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6 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. By backing out a contribution share, this paper contributes to our understanding of which team structures promote better innovations. I demonstrate that as inventors become increasingly experienced, they collaborate on more patents each year. However when they are relatively more senior than their co-inventors, they tend to contribute less to each patent. This can have important consequences for patent value, as I corroborate the literature in showing that concentrating contribution shares onto one or a few team members tends to decrease patent value.

How teams organise themselves to produce knowledge has traditionally been a black-box, and the model presented in this paper allows us to disentangle their contributions given their observed output. Beyond shedding light on within-team dynamics, this model presents a platform for future work examining how teams collaborate to develop the scientific and technological frontier. I believe the following are first order questions within the organisation of the production of science and technology that this paper will allow us to explore.

Future research should link contribution measures with wage data to test whether innovation is fairly rewarded. Are junior and senior inventors compensated in line with the significance of their input, and which contributions truly constitute the "special sauce"? It will also be important to connect these measures with occupational and hierarchical data. Do formal structures of authority shape idea generation, and can we directly assess the role of managers in driving innovation? Finally, this approach offers a way to uncover the contributions of under-represented groups in science. Tracing the impact of women, ethnic minorities, and other marginalised groups, and asking whether their work is recognised and rewarded—will be essential to ensuring that technological progress serves society at large.

References

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

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.

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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.

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.

Jones, **Charles I.** (2023). "Recipes and Economic Growth: A Combinatorial March Down an Exponential Tail". *Journal of Political Economy* 131.8, pp. 1994–2031.

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 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.

Marx, **Matt** and **Aaron Fuegi** (2022). "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations". *Journal of Economics & Management Strategy* 31.2, pp. 369–392.

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.

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

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

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.

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.

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.

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

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

Appendix

A Inferring the Knowledge Space: Intuition

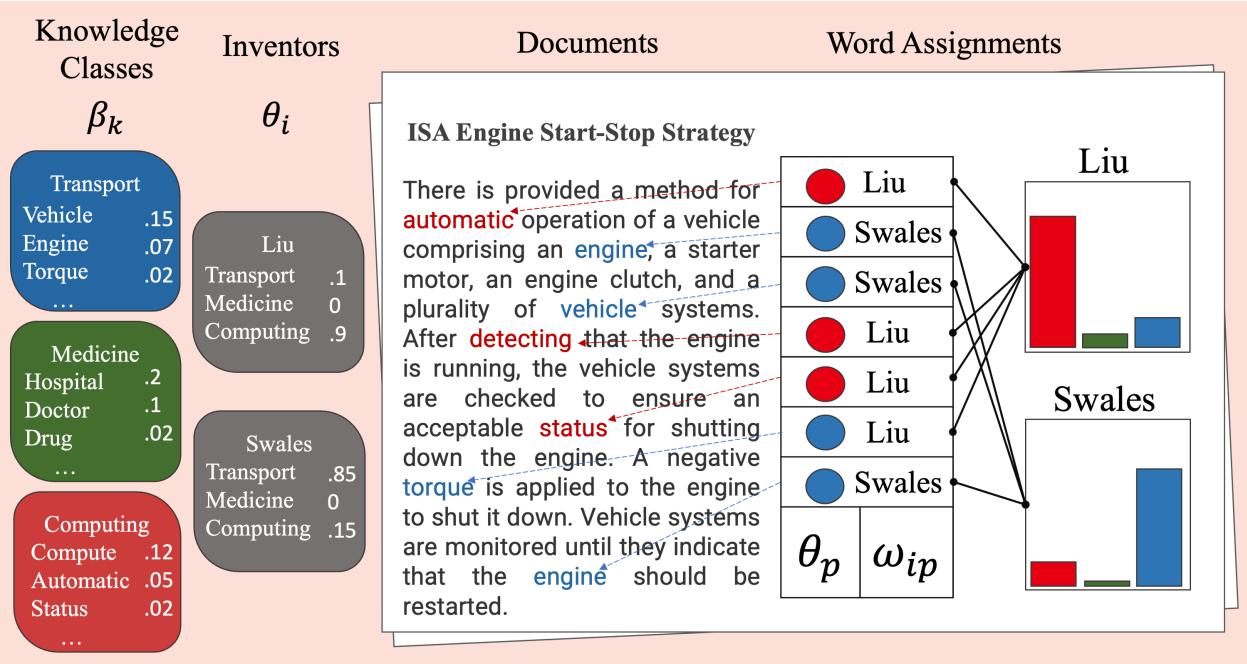
Figure A1 provides an intuitive example of how the Author-Topic model learns the knowledge classes, inventor knowledge profiles and can describe the knowledge content of a patent and an inventor's contribution share. The knowledge classes (β_k) and inventor knowledge distributions (θ_i) are latent variables. 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 A1

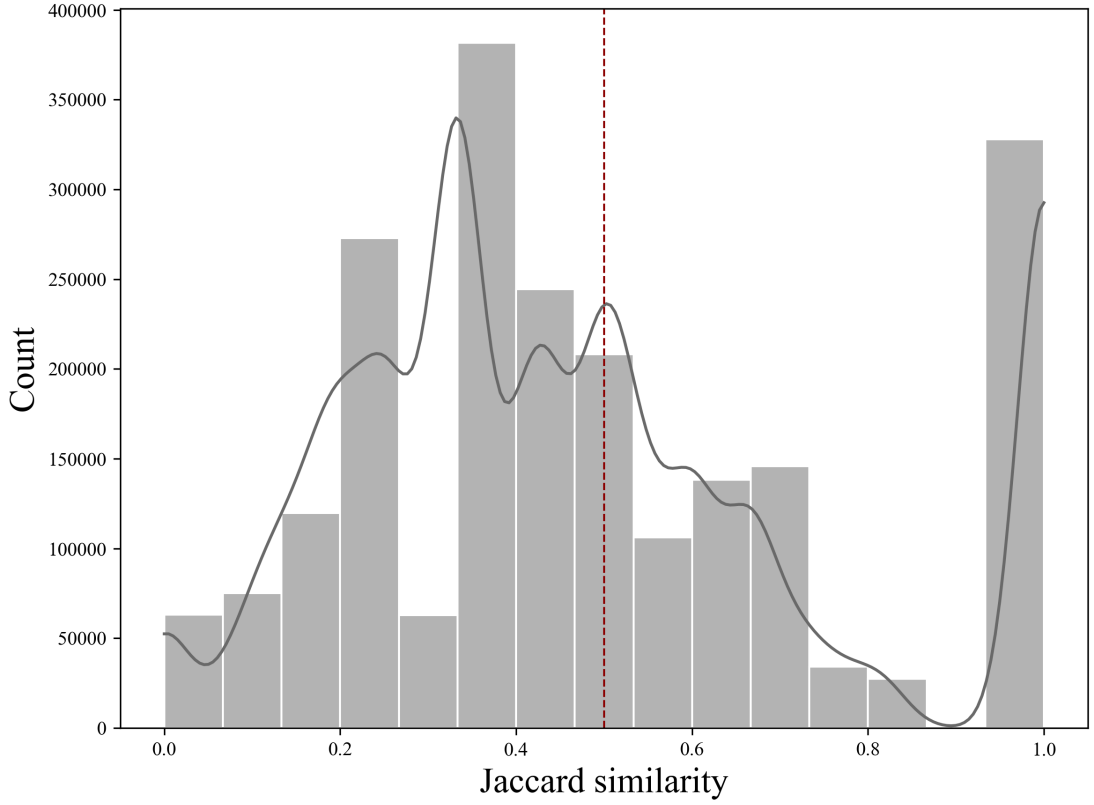
INTUITIVE LDA EXAMPLE



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.

B Additional Tables and Figures

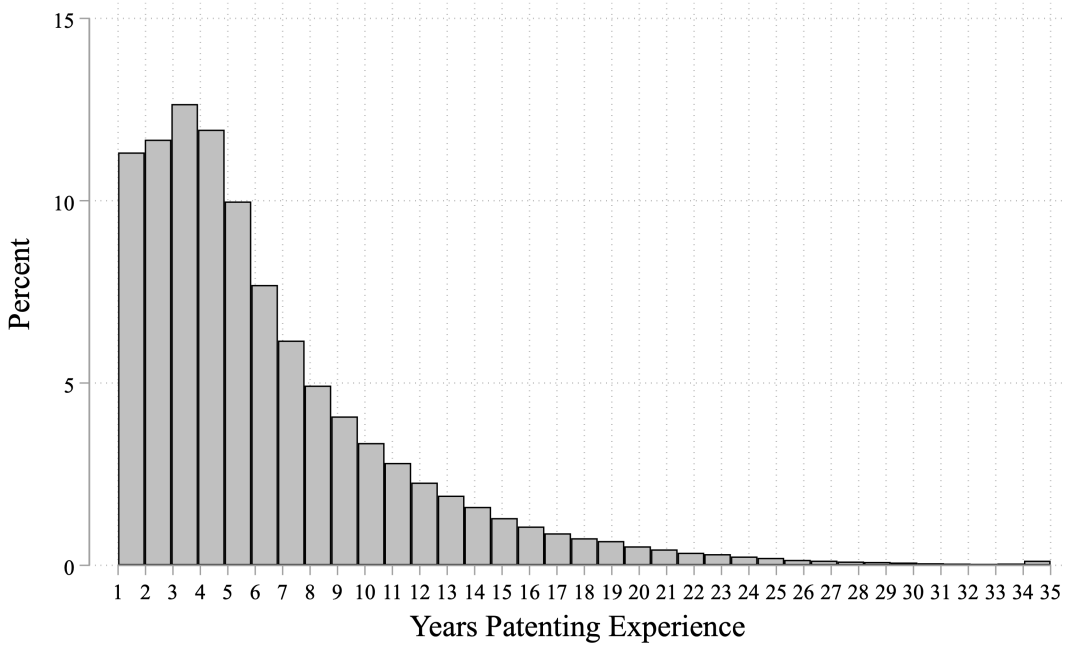
Figure B2
HISTOGRAM OF TEAM JACCARD SIMILARITIES



Notes: A histogram of the Jaccard similarity between the lead and last inventor on a team patent. The lead is defined as the inventor with the largest contribution share, the last is the lowest share on that patent. The Jaccard similarity is defined as:

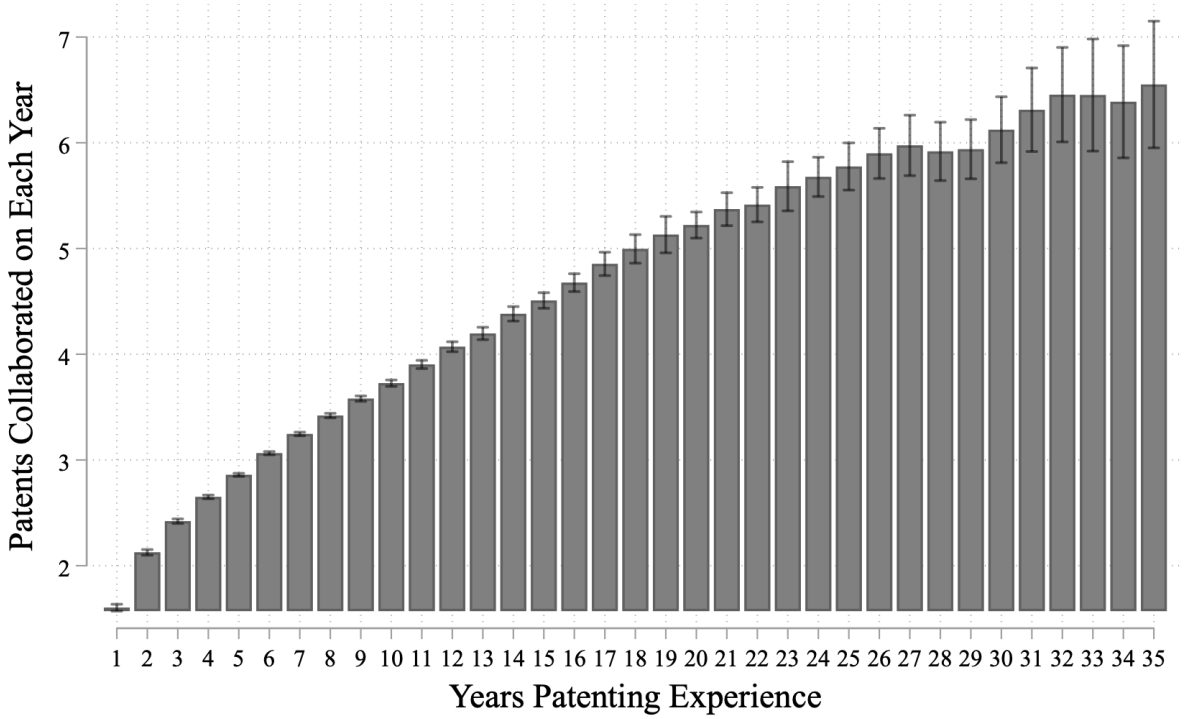
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Figure B3
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. The sample corresponds to the set of inventors, belonging to the giant component of the inventor network for collaborations from 1976-2024, who have more than 5 patents.

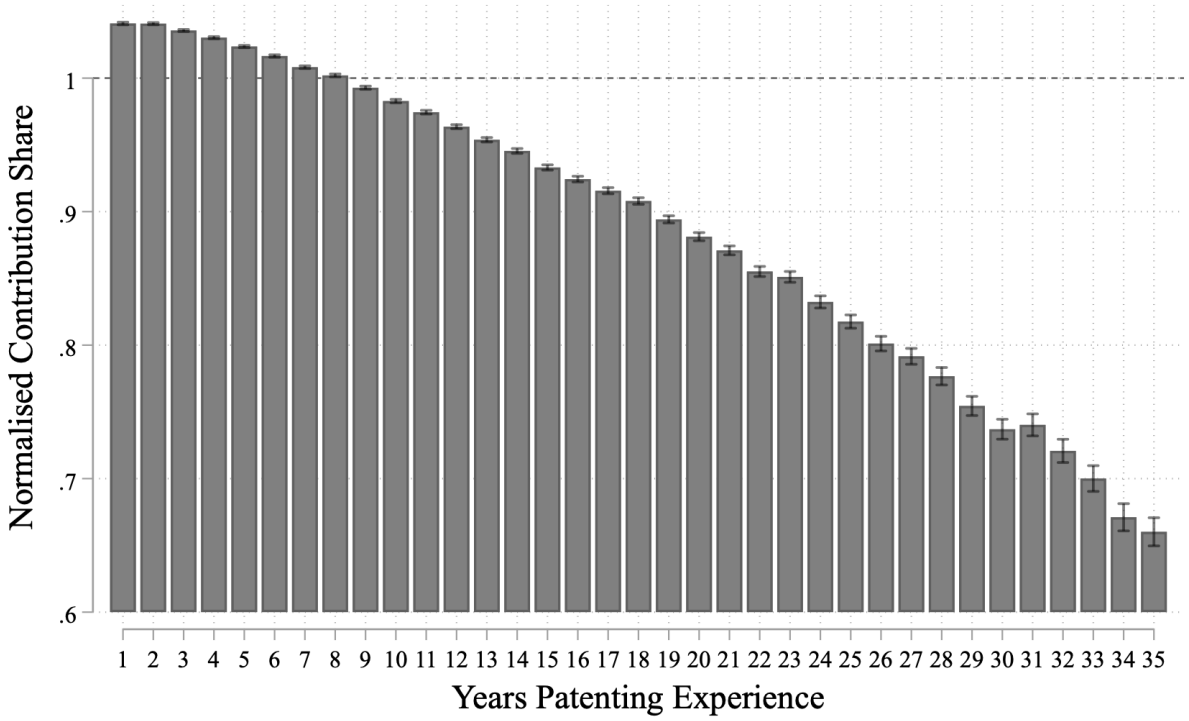
Figure B4
PATENTS PER YEAR OF EXPERIENCE



Notes: The average number of patents that an inventor collaborates on, in each year of their patenting career. Years of patenting experience refers to the number of individual years in which they appear on a patent. The average value is taken from the predicted values of a regression of the number of patents an inventor collaborates on in one year, an inventor fixed effect, a year fixed effect and set of year experience dummies. N = 3,262,327.

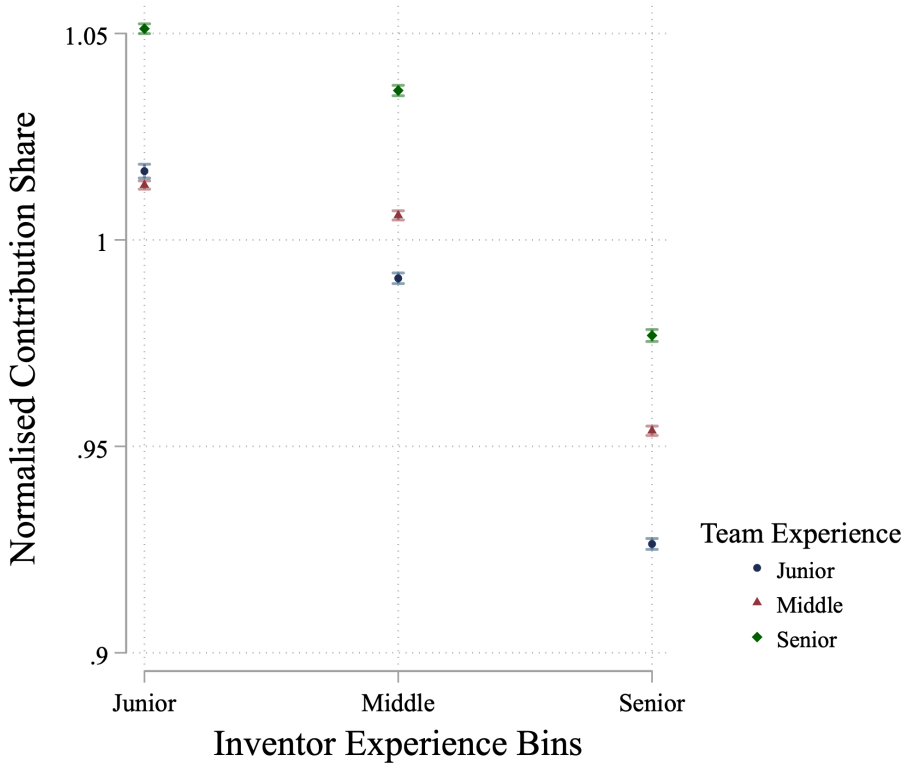
Figure B5

CONTRIBUTION SHARE PER YEAR OF EXPERIENCE: WITHIN TEAM



Notes: Plots the average normalised contribution share for inventors in each year of their patenting career. Years of patenting experience refers to the number of individual years in which they appear on a patent. The normalised share is defined as their contribution share, weighted by the inverse of team size. A normalised share equal to one means that they are contributing a share that would be expected within an evenly distributed team. The predicted values are taken from a regression of the normalised share on a set of patent fixed effects, and year experience dummies. Standard errors are clustered at the patent level. N = 6,931,882.

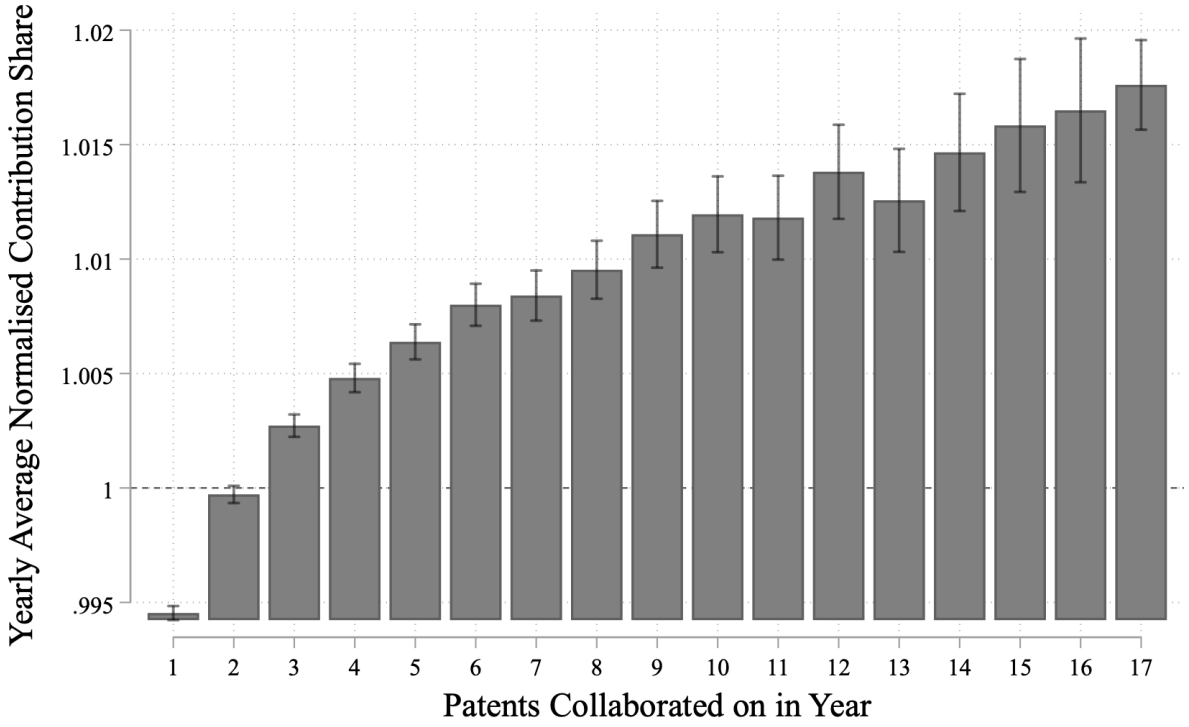
Figure B6
CONTRIBUTION SHARE BY RELATIVE SENIORITY: WITHIN TEAM



Notes: Plots the average contribution share for each of the nine combinations of inventor by team experience levels. Juniors have fewer than 4 years experience, Middle have 5-8 and Seniors have between 8-35 years of experience. Team experience refers to the average years patenting experience in the team, minus the target inventor, again split into terciles. The normalised share is defined as their contribution share, weighted by the inverse of team size. The predicted values are taken from a regression of the normalised share on a set of patent fixed effects, and dummy variables for the inventors experience bin, the average experience of their team and the interaction. Standard errors are clustered at the patent level. N = 6,931,882.

Figure B7

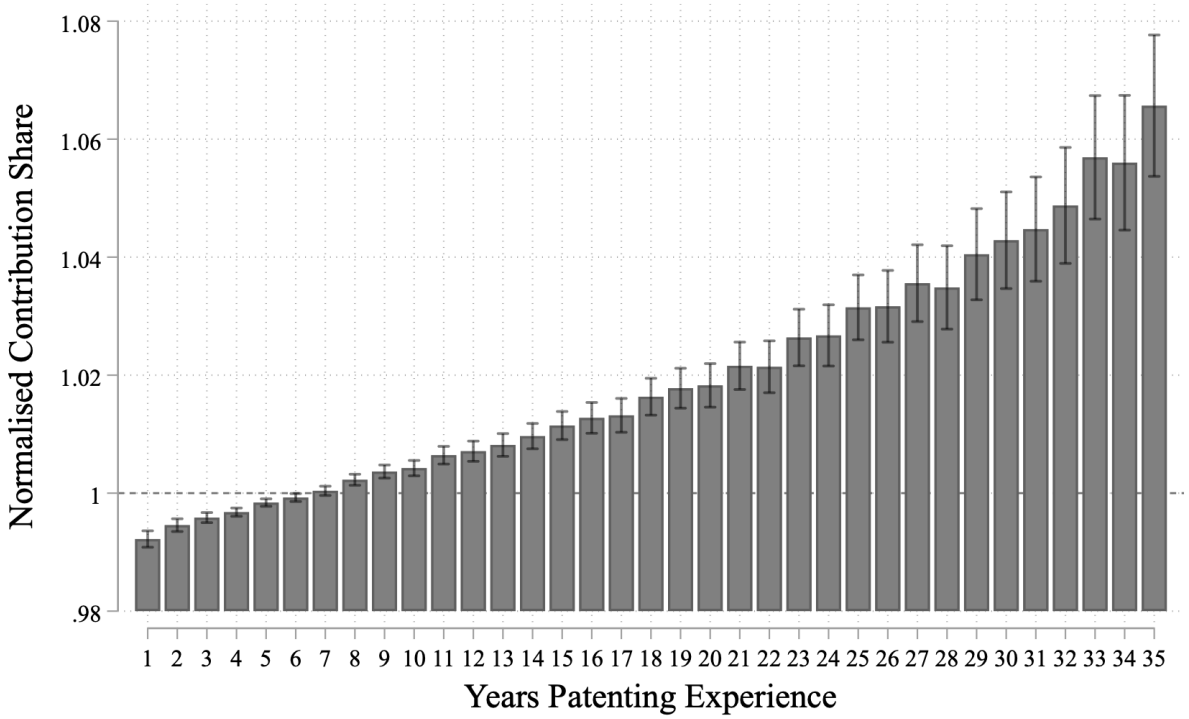
CONTRIBUTION SHARE PER YEARLY PATENT COUNT: WITHIN INVENTOR



Notes: Plots the yearly average normalised contribution share for an inventor across the number of patents they collaborate on in that year. The normalised share is defined as their contribution share, weighted by the inverse of team size. The predicted values are taken from a regression model. The dependent variable is the average normalised contribution share for an inventor in a given year. This is regressed on a set of inventor fixed effects, year patenting dummies and a dummy for the number of patents they collaborate on in that year. Standard errors are clustered at the inventor level. N = 3,262,327.

Figure B8

CONTRIBUTION SHARE PER YEAR OF EXPERIENCE: WITHIN INVENTOR



Notes: Plots the average normalised contribution share for inventors in each year of their patenting career. The normalised share is defined as their contribution share, weighted by the inverse of team size. A normalised share equal to one means that they are contributing a share that would be expected within an evenly distributed team. The predicted values are taken from a regression of the normalised share on a set of inventor fixed effects, and year patenting dummies. Standard errors are clustered at the inventor level. N = 3,262,327.