Title: Inventing modern invention: the professionalization of technological progress in the US

Authors: Matte Hartog[1*], Andres Gomez-Lievano[1,3], Sultan Orazbayev[1], Ricardo Hausmann[1], Frank Neffke[1,2]

Affiliations: [1]Growth Lab, Harvard University; 79 JFK Street, Cambridge, MA 02138, USA
[2]Complexity Science Hub; Josefstädter Str. 39, 1080 Wien, Austria
[3]Analysis Group; 111 Huntington Ave 14th floor, Boston, MA 02199, USA

[*]To whom correspondence should be addressed; E-mail: matte_hartog@hks.harvard.edu

Abstract: The period between 1920 and 1950 witnessed a remarkable surge in novel technological combinations on US patents. After digitizing 245,604 pages of information on inventors, organizations and technologies in over 1.6M patents, and integrating this with the full US census from 1850 to 1940, we show that this exponential rise in novel combinations was not driven by particular technological advances, but part of a wider, transformative revolution in the organization of invention across all sectors. Specifically, we posit that there was a paradigm shift in inventive activity and we introduce the concept of "System 2" to refer to a new way of inventing that replaced the traditional "System 1." System 2 emerged through the professionalization of engineers and the replacement of collaborations within families with professional ties within firms and industrial research labs. This revolution not only altered the division of labor but also reshaped the geography of innovation, with firms and research labs facilitating long-distance collaboration and reestablishing large cities as epicenters of inventive activity, and implied new barriers to participation for women and foreign-born inventors.

One-Sentence Summary: Social arrangements, evolving from family to industrial labs, catalyzed the surge of radical innovations in the US.

Main Text: As Joseph Schumpeter proposed, new technologies often arise from combining existing technologies. As the repertoire of technologies expands, so does the potential

for novel combinations [1, 2, 3], which in principle should make further innovation easier. However, research has shown that navigating this exponentially expanding technological landscape poses challenges such as diminished research productivity [4], longer learning curves [5] and greater interdependencies among experts [6]. The prevailing focus on characteristics of technology or inventors, however, leaves unexplored *how* knowledge is combined. We argue that combining human expertise does not happen spontaneously and that societies need more than technological know-how to innovate: they must find stable social arrangements to coordinate a growing body of distributed knowledge. So far little systematic investigation exists into the means by which this coordination is achieved.

To study the relationship between innovation and knowledge coordination, we follow the US economy from its agricultural beginnings to its establishment as global leader in technology, science and industry at the cusp of World War II. For this period, we digitize historical records of 1.6M patents of US inventors granted by the US Patent and Trademark Office (USPTO) and merge them with the US census and with information on industrial research labs. We supplement this information with existing datasets on 1.8M patents in later years. These data allow us to describe the evolution of US invention and its novel combinations over a much longer time period than prevailing studies [7, 8, 9, 10].

To identify the factors that drove the generation of novelty during this period, we determine if the technological fields mentioned in a patent had already been combined in previous patents. Applying this procedure at the level of the 474 3-digit technology codes on the one hand, and at the complete set of over 150k 6-digit codes on the other, allows us to distinguish between "radical" innovation – new combinations of broad technology classes – and "incremental" innovation – combinations that are new in terms of detailed, but not broad technology classes.

We find that the rapid increase in radically novel technological combinations early in the twentieth century developed in concert with a shift in *how* the US invented. Labor inputs fueling invention changed and collaborations within families were replaced by professional ties within firms and, later, within industrial research labs. We regard this shift in the early 20th century in the locus of coordination from families to labs as an important social innovation in how invention was done in the US. We will refer to the old way of inventing as "system 1" and to the new way as "system 2". The emergence of system 2 manifested itself not only in increased novelty, but also in a changing division of labor and a changing geography of innovation, where firms and industrial research labs facilitated long-distance coordination and helped reestablish the primacy of large cities as the main locus of invention.

These results support the perspective that with increasingly complex technologies, societies must introduce new social arrangements to coordinate human creativity and expertise. Thus, it suggests social change does not simply follow technological change, but rather, it also enables it, consistent with arguments put forward by anthropologists who have studied technological change [e.g., 11, 12, 13].

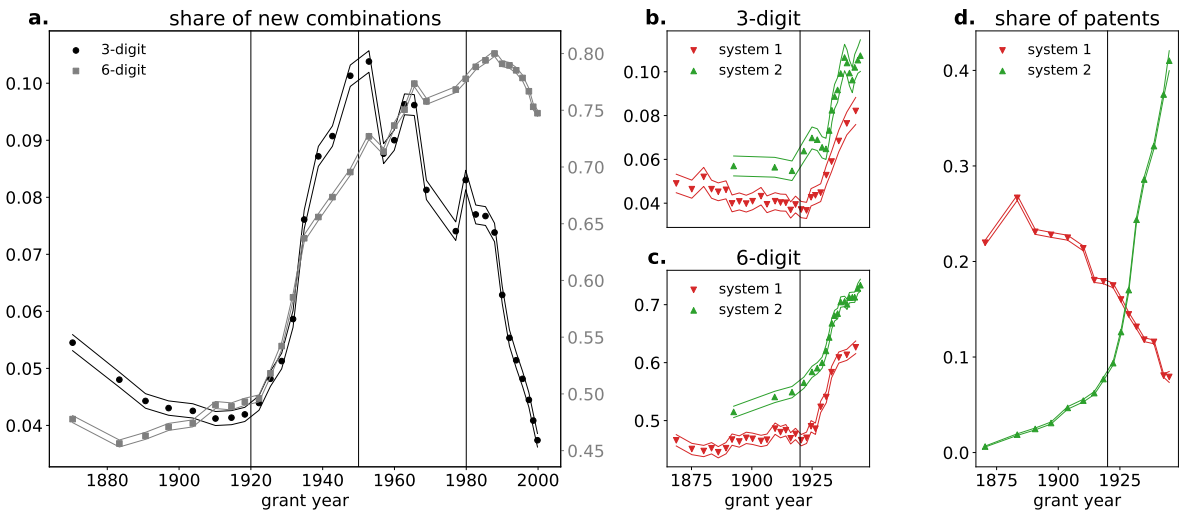


Figure 1: Nature of invention. **a**: Share of combinations of 3- (black) or 6-digit (gray) technology classes that had never been used before. **b**/**c**: idem a, but now for inventions produced in system 1 (red) and system 2 (green). **d**: Share of patents produced by system 1 (red) and system 2 (green). Note that a third group of patents that does not clearly belong to one or the other system has been omitted here.

Results

Fig. 1 reveals an explosion in new technological combinations on US patents roughly between 1920 and 1950, a period known as the golden age of American ingenuity [14]. This rapid rise in novel combinations is not limited to a specific set of technologies (SI, Fig. 17). Instead, it suggests a revolution in the organization of invention that fundamentally changed how inventions were produced across all sectors. This includes shifts in the demographic profile of inventors, as well as increases in teamwork, which have so far only been documented for the late 20th century (e.g., [15]). With our data, we pinpoint the beginnings of this process and probe deeper into concurrent changes in labor inputs, the nature and coordination of these teams, the geography of system 1 and system 2, and the exclusion of some groups in society from these processes.

Labor inputs

In the 1920s, labor inputs into US invention start changing drastically (Fig. 2). First, we witness the rise of a new type of inventor: the engineer. Whereas 19th century invention was dominated by blue collar workers and mechanics, in the 20th century, engineers start dominating invention to a hitherto unseen extent: in the 1940s, while engineers accounted for only 0.7% of the US labor force, the share of engineers among inventors reaches 25%

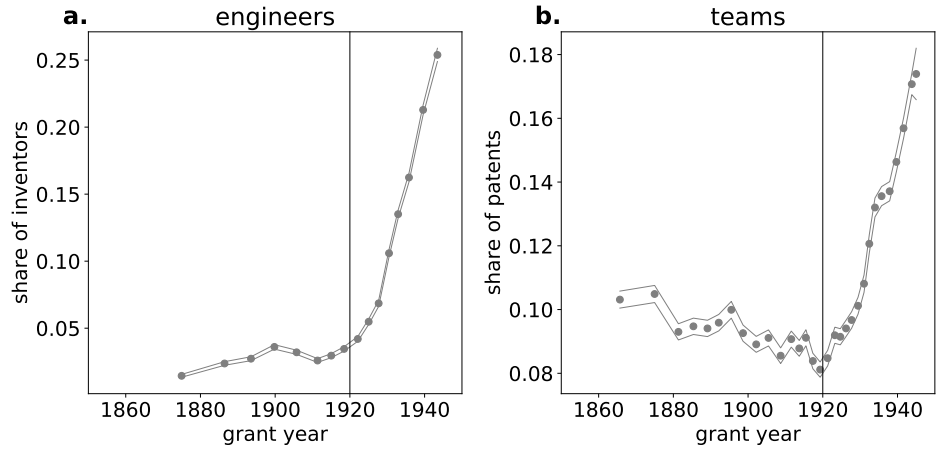


Figure 2: Labor inputs in invention. **a**: Share of inventors who are engineers. **b**: share of patents filed by teams. Lines display 95% confidence intervals.

(see SI, Fig. 7a). With the rise of engineers, we also witness the fledgling start of patenting by academics (see SI Fig. 20). In fact, because most academic patents were assigned to firms, they represented early examples of university-industry collaborations. However, it would take another 20 years for universities to start claiming ownership of such patents. Second, we see that inventors start working more and more in teams. In fact, the well-known rise of teamwork in invention from the 1970s onwards [15] had an abrupt start in the 1920s, when a slow but steady decline suddenly turned into a rapid increase of the share of patents created by teams. In SI section A.4.1, we show that both engineers and teams are behind a disproportionate share of novel technological combinations.

Team coordination

This rise of teamwork coincides with a change in how teams are coordinated. In particular, coordination shifts from family ties to professional ties (Fig. 3): whereas in 19th century teams, inventors were often related – as inferred from the fact that they shared the same last name – 20th century teams often worked for firms and, from the 1920s on, in industrial research labs. In fact, teamwork was particularly common in research labs: in the 1940s the share of teams on research lab patents is twice as large as on "stand-alone" patents - those not assigned to labs or firms (Fig. 3b).

Teams in firms, and even more so in research labs, were significantly more likely to introduce novel combinations than stand-alone teams (Fig. 3d). Our data suggest that labs and firms conferred certain advantages for this to their teams. First, they seem to have created stable environments that fostered repeated collaboration among inventors (Fig. 3c). Second, teams in firms and research labs were able to work together over longer

distances and their team members were typically closer in age and more likely to share the same occupations (see Fig. 19 of the SI).

System 1 and system 2 invention

Taken together, these findings suggest that the shift to engineers and firm-based and later lab-based teamwork played an important role in the rise of combinatorial exploration that started in the 1920s. To describe the spatial and societal consequences of this shift, we define two archetypal systems of invention. *System 1* dominates in the 19th century and consists of patents by solo inventors outside firms and without engineering backgrounds. *System 2* takes off in the 1920s and consists of patents that are either invented by engineers, or created in research labs or by firm-based teams.

These definitions to some extent mimic the Schumpeter Mark I and Mark II innovation processes distinguished by [16]. However, whereas Schumpeter Mark I innovation is typically associated with creative destruction and radical change and Mark II innovation with creative accumulation and incremental change [17], our findings in Fig. 1b suggest the opposite: it is the science-based invention organized in firms and labs that introduced radically new combinations of technologies, not the artisanal invention by individual inventors of system 1.

However, the link between radical innovation and the firm-based teamwork of system 2 weakens after 1950 and reverses in the last quarter of the 20th century, when stand-alone teams more often create new combinations than firm-based teams. In the SI, we show that this finding is robust to controlling for technology-specific time trends. However, it only holds for radical, 3-digit, not for incremental, 6-digit combinations. This suggests that only in the second halve of the 20th century, system 2 starts shifting to the incremental invention that is typically associated with Schumpeter Mark II innovation.

Geography

System 1 and 2 also exhibit divergent spatial patterns. First, relative to the population's spreading out across the US, over the course of the 19th century, patenting became more spatially concentrated (Fig. 4a). However, in the 20th century, this process reverses and invention starts diffusing across more and more cities. This reversal first happens in system 2 and then somewhat later in system 1 as well. In parallel, large cities first lose in importance and then start regaining their mid-19th century primacy (Fig. 4b). This suggests that, contrary to findings in [18], there is no continuously strengthening concentration of invention in large cities. Instead, there is a shift from large cities first falling and then rising in patenting. Again, the reversal is led by system 2, with system 1 following some years later. This leader-laggard dynamic is also visible in Fig. 4c, which shows that the spatial profiles of system 1 and 2 diverge until 1920, after which they rapidly converge. Ultimately, the geographical patterns of system 2 come to describe the

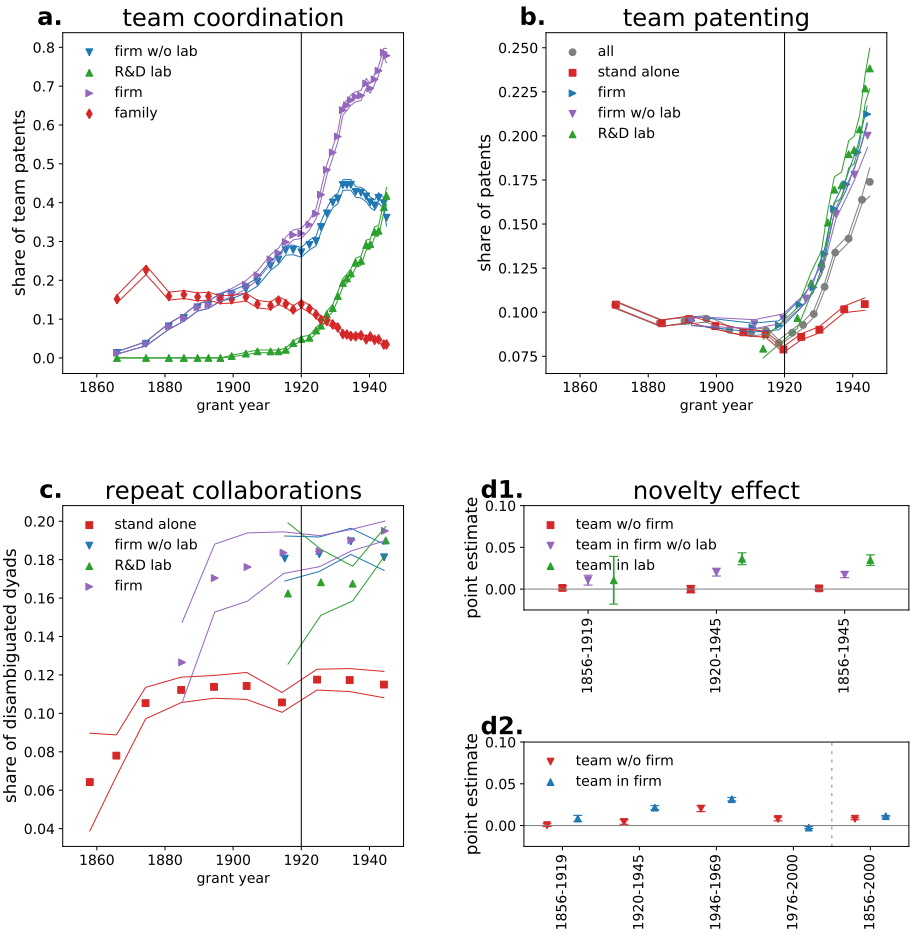


Figure 3: Team coordination. **a**: Share of team patents that are coordinated by firms without labs, firms with labs or families. **b**: Share of patents that are team patents by coordination mechanism. **c**: Stability of inventor dyads: share of patents resulting from repeated collaborations (i.e., by inventor pairs that patent more than once together in the 10-year period). This is calculated as a share of patents by dyads that can be reliably disambiguated because inventor names are relatively rare (see methods). **d**: increase in share of new combinations of 3-digit technologies by team type from a linear regression model (see methods). Upper panel: 1856-1945 (including labs); lower panel: 1856-2000 (only firms). Purple: all firms, blue: firms without industrial research labs, green: firms with industrial research labs, red: families (at least two inventors on patent share the same last name). Lines display 95% confidence intervals.

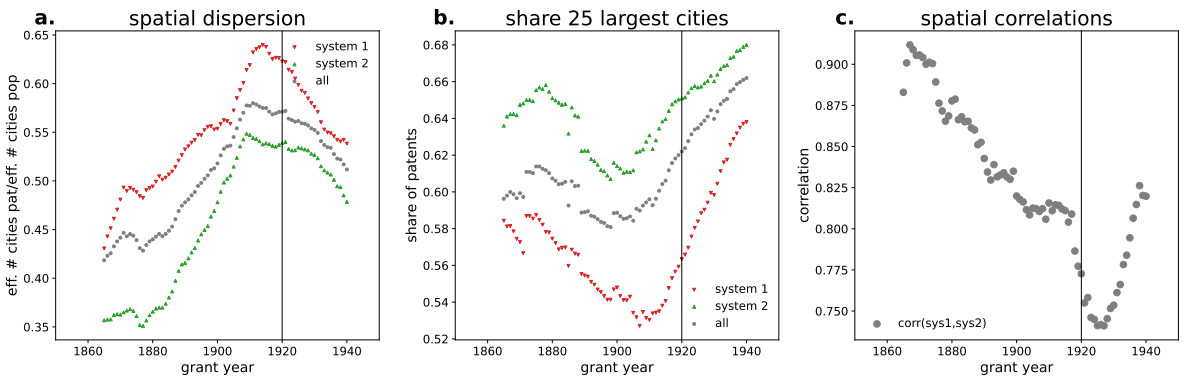


Figure 4: Geography. **a**: Concentration of inventive activity: effective number of cities – defined as the e^H, where H is the entropy of the distribution of patents across cities (see methods) – divided by the effective number of cities in terms of population **b**: Share of patents in the 25 most populous cities in the US. **c**: Correlation between vectors that reflect the overrepresentation of patenting across cities vis-à-vis their population. Gray: all patents, red: system 1 patents, green: system 2 patents.

rise of an area that is nowadays known as the American "rustbelt" (see SI Fig. 18), but at the time represented the cutting edge of the US economy.

Participation

System 1 and 2 are associated with different inputs, coordination mechanisms, geographical patterns and novelty. The two systems also differ in terms of participation by women and foreign-born inventors (Fig. 5). Patenting by women has steadily increased over the course of the 20th century, peaking immediately after World Wars I and II. However, this increase and these peaks are only visible in the non-firm patenting of system 1. In firms, female inventor shares were about a factor 5, in research labs a factor 7, lower than among stand-alone inventors. This gap only starts closing in the late 1970s, but even in the year 2000, the share of female inventors is about 25% lower in firms than among stand-alone inventors. Gender gaps may exist due to sparser connections in women's social networks to industry [19] or to successful mentors [20] and role models [21], differences in patentability in fields where women are most active [19] or outright biases that exclude them from being listed as coinventors [22]. Although we cannot determine the deeper causes of gender gaps in patenting, our analysis points to another possible explanation: differential access to research-lab-based and firm-based invention.

When it comes to national origins, foreign-born inventors have played an important role in US invention [14]. In the 1920s, the share of foreign-born inventors rapidly decreases across the board, but contributions by foreign-born inventors are dropping faster

in system 2, where they, like women, are generally underrepresented. However, this differs across nationalities. For instance, whereas inventors with Hispanic surnames are underrepresented, inventors with East-Asian surnames are over-represented in firm-based invention (Fig. 5d).

Discussion

Our study reveals a pivotal moment in the history of US invention. In particular, the 1920s' rapid rise in new technological combinations goes hand-in-hand with the adoption of an organizational innovation: the industrial research lab. These labs seem to offer certain advantages. First, they lead the trend in hiring a new class of specialist inventors: engineers. Second, they help coordinate expertise across inventors, allowing for more frequent teamwork, more stable teams and long-distance collaborations. This organizational innovation diffuses across cities and becomes rapidly adopted by the most prolifically innovating firms: between 1940 and 1945, out of a total of almost 15,000 different patenting firms, 25% of all patents and 42% of all firm-based patents in the US are created by just 1,172 firms that operated research labs. This probably represents an under-count: many firms may have adopted research labs or similar organizational units, without being listed in the survey. Moreover, of the 8,000 unique lab names that are reported in the surveys, we could match only 1,994 to patent assignees.

However, from the 1950s on, as research labs and the teams they organize become widely adopted, the capacity of firms to generate radical novelty declines. This resonates with the finding in [23] that the very large teams we see today struggle to propose radically new ideas in science and technology. One challenge of running large teams is that the number of bilateral links that need to be coordinated increases with the square of the number of team members. Based on the arguments put forward in this paper, a potential explanation for the reduced capacity of teams to create radically new technological combinations is that existing organizational forms cannot meet the increased coordination demands in increasingly large teams. This raises the question of whether accelerating technological innovation requires another round of organizational innovation and whether new online collaboration technologies, such as Slack, Zoom and similar platforms can play this role to help overcome current organizational bottlenecks in teamwork and set in motion a new wave of radical technological change.

Materials and methods summary

Our analyses combine three different types of data: about half a million scanned pages of patent yearbooks, the full US Census from 1850 to 1940 with almost a billion records on individuals (except 1890 when a fire destroyed the Census records), and large-scale

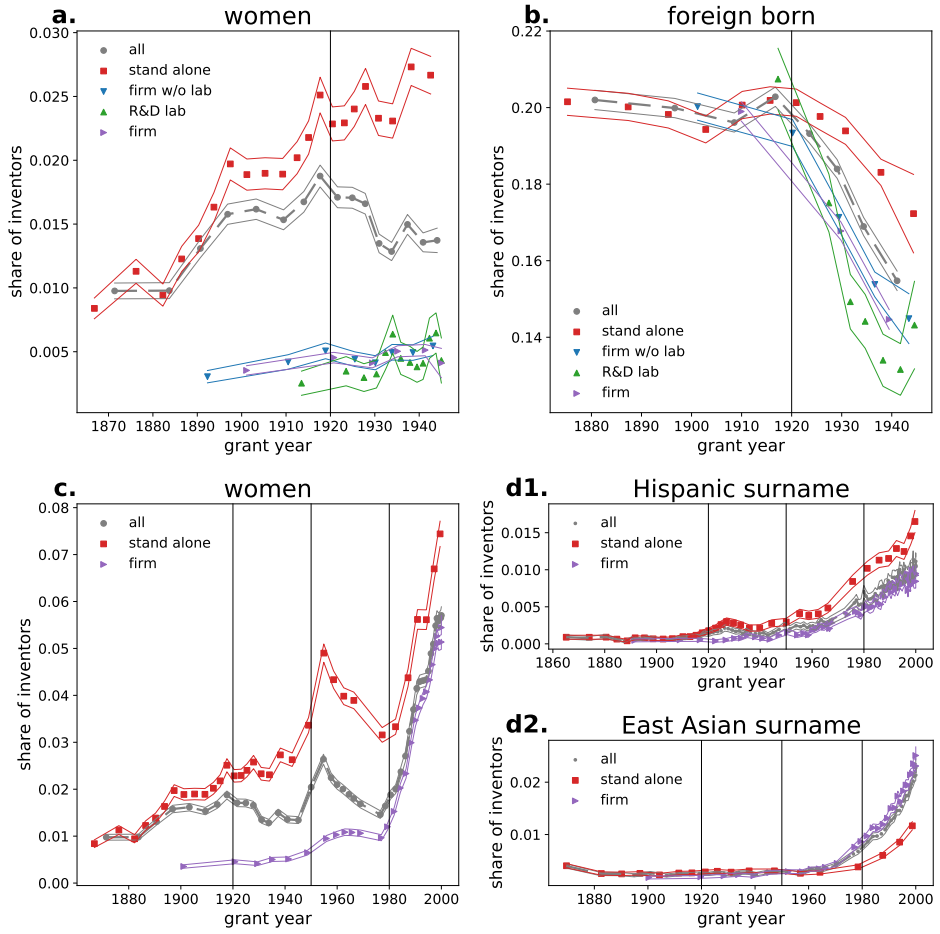


Figure 5: Participation in system 2 by gender and ethnicity. **a**: share of inventors with first names that are predominantly used by women in system 1 and system 2 invention (1856-1945) **b**: share of foreign-born inventors in system 1 and system 2 (1856-1945). **c**: share of inventors with first names that are predominantly used by women in system 1 and system 2 invention patents (1856-2000). **d** upper panel: share of inventors with Hispanic surname, lower panel: share of inventors with East-Asian surnames (see methods).

surveys of industrial research labs in the US between 1920 and 1956. We digitized information related to the patents and industrial research labs using optical character recognition and natural language processing, and then matched inventors and assignees across data sources. The final data base consists of a unified longitudinal panel of patents and their technologies, inventors and their demographic characteristics (including occupation and location) and organizations and their research labs. Finally, these data sets were supplemented with existing digitized patent records from PATSTAT (1953-1969) and PatentsView (1975-2000). The entire process is summarized in Fig. 6 of the SI.

We group inventors' occupations into higher level aggregates based on an analysis of clusters in a cross-occupational labor flow network derived from matched census records. We focus on a cluster that we call "engineers", which encompasses all engineering occupations, as well as some other closely related occupations with strong labor-flow connections to them. This alleviates concerns that the observed rise of engineers merely reflects a change in terminology or occupational identity. Fig. 7 of the SI compares the overrepresentation of broad occupational groupings among inventors using the flow-based clusters and traditional groupings, corroborating the outsized role of engineers.

We identify family collaborations as patents with at least two inventors who share the same last name. In the period 1856-1945, the likelihood of this happening by chance is negligible (SI Fig. 8). Although this procedure only captures a subset of family-linkages, it allows to identify these links on any patent, not just for inventors whom we could match to the census data.

For similar reasons, we infer inventors' gender from their first names. To do so, we assess how often a first name was associated with a given gender in the 600M census records. Next, when analyzing gender effects, we focus on a sample for which the link between first name and gender is not ambiguous, requiring that at least 90% of census records agree on one gender. As a robustness check, Fig. 9 of the SI repeats the analyses of Fig. 5 for the more ambiguous records we dropped, associating each first name with the modal gender in the census. We follow a similar procedure to identify Hispanic and East-Asian last names. However, because these last names were less common in the US before 1945, we rely on existing pre-trained algorithms (sec. A.2.6).

References

[1] Martin L Weitzman. Recombinant growth. *The Quarterly Journal of Economics*, 113(2):331–360, 1998.

[2] W Brian Arthur. *The nature of technology: What it is and how it evolves*. Simon and Schuster, 2009.

[3] Hyejin Youn, Deborah Strumsky, Luis MA Bettencourt, and José Lobo. Invention

as a combinatorial process: evidence from us patents. *Journal of the Royal Society interface*, 12(106):20150272, 2015.

[4] Nicholas Bloom, Charles I Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020.

[5] Benjamin F Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

[6] Frank MH Neffke. The value of complementary co-workers. *Science advances*, 5(12): eaax3370, 2019.

[7] Lee Fleming. Recombinant uncertainty in technological search. *Management science*, 47(1):117–132, 2001.

[8] Deborah Strumsky and José Lobo. Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8):1445–1461, 2015.

[9] Dennis Verhoeven, Jurriën Bakker, and Reinhilde Veugelers. Measuring technological novelty with patent-based indicators. *Research policy*, 45(3):707–723, 2016.

[10] Michele Pezzoni, Reinhilde Veugelers, and Fabiana Visentin. How fast is this novel technology going to be a hit? antecedents predicting follow-on inventions. *Research Policy*, 51(3):104454, 2022.

[11] Stephen Shennan. Demography and cultural innovation: a model and its implications for the emergence of modern human culture. *Cambridge archaeological journal*, 11(1):5, 2001. URL `http://search.proquest.com/openview/4310d17e353f71d682b6baefe82a6c3f/1?pq-origsite=gscholar&cbl=13203`.

[12] Joseph Henrich. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the tasmanian case. *American Antiquity*, 69(2): 197–214, 2004.

[13] Michael Muthukrishna and Joseph Henrich. Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690): 20150192, 2016.

[14] Ufuk Akcigit, John Grigsby, and Tom Nicholas. Immigration and the rise of american ingenuity. *American Economic Review*, 107(5):327–331, 2017.

[15] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827):1036–1039, May 2007.

[16] Richard R Nelson. *An evolutionary theory of economic change*. harvard university press, 1982.

[17] Stefano Breschi, Franco Malerba, and Luigi Orsenigo. Technological regimes and schumpeterian patterns of innovation. *The economic journal*, 110(463):388–410, 2000.

[18] Pierre-Alexandre Balland, Cristian Jara-Figueroa, Sergio G Petralia, Mathieu PA Steijn, David L Rigby, and César A Hidalgo. Complex economic activities concentrate in large cities. *Nature human behaviour*, 4(3):248–254, 2020.

[19] Waverly W Ding, Fiona Murray, and Toby E Stuart. Gender differences in patenting in the academic life sciences. *science*, 313(5787):665–667, 2006.

[20] Mercedes Delgado and Fiona Murray. Faculty as catalysts for new inventors: Differential outcomes for male and female phd students. In *Academy of Management Proceedings*, number 1, page 17297. Academy of Management Briarcliff Manor, NY 10510, 2022.

[21] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly Journal of Economics*, 134(2):647–713, 2019.

[22] Matthew B. Ross, Britta M. Glennon, Raviv Murciano-Goroff, Enrico G. Berkes, Bruce A. Weinberg, and Julia I. Lane. Women are credited less in science than men. *Nature*, 608(7921):135–145, August 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04966-w.

[23] Lingfei Wu, Dashun Wang, and James A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, February 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-0941-9.

[24] Gaurav Sood and Suriyan Laohaprapanon. Predicting race and ethnicity from the sequence of characters in a name, 2018.

[25] Anurag Ambekar, Charles B. Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. Name-ethnicity classification from open sources. In *Knowledge Discovery and Data Mining*, 2009.

[26] European Patent Office. Patstat - worldwide patent statistical database - spring 2020, 2020.

[27] USPTO. PatentsView. `https://www.patentsview.org/download/`, 2022.

[28] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. *IPUMS USA: Version 11.0 [Dataset]. Minneapolis.* IPUMS Minnsesota: MN, January 2021.

[29] National Research Council. *Industrial Research Laboratories of the United States, Including Consulting Research Laboratories.* National Research Council, Washington, D.C., 1956.

[30] Jeffrey L. Furman and Megan J. MacGarvie. Academic science and the birth of industrial research laboratories in the U.S. pharmaceutical industry. *Journal of Economic Behavior & Organization*, 63(4):756–776, August 2007. ISSN 01672681. doi: 10.1016/j.jebo.2006.05.014.

[31] Alan C. Marco, Michael Carley, Steven Jackson, and Amanda F. Myers. The uspto historical patent data files: Two centuries of innovation. *SSRN working paper*, June 2015. URL `http://ssrn.com/abstract=2616724`.

[32] Bronwyn H Hall, Adam B Jaffe, and Manuel Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools, 2001.

[33] Frank Neffke and Martin Henning. Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3):297–316, 2013.

[34] Alje van Dam, Andres Gomez-Lievano, Frank Neffke, and Koen Frenken. An information-theoretic approach to the analysis of location and co-location patterns. *arXiv preprint arXiv:2004.10548*, 2020.

[35] Rachata Muneepeerakul, José Lobo, Shade T Shutters, Andrés Goméz-Liévano, and Murad R Qubbaj. Urban economies and occupation space: Can they get "there" from "here"? *PloS one*, 8(9):e73676, 2013.

[36] Yang Li and Frank Neffke. Relatedness in regional development: in search of the right specification. *arXiv preprint arXiv:2205.02942*, 2022.

[37] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[38] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[39] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2009.

[40] Bronwyn H. Hall, Adam B. Jaffe, and Manuel Trajtenberg. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Working Paper 8498, National Bureau of Economic Research, October 2001.

[41] Jan Tinbergen. *Shaping the world economy; suggestions for an international economic policy*. Twentieth Century Fund, 1962.

[42] JMC Santos Silva and Silvana Tenreyro. The log of gravity. *The Review of Economics and statistics*, 88(4):641–658, 2006.

Acknowledgements

Acknowledgements: The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. We appreciate S. Petralia for introducing us to the patent books and A. Grisanti for helping us create ground truths for optical character recognition purposes.

A Materials and methods

A.1 Data sources

Our analyses combine three different types of data. The main data set contains information on patents issued by the United States Patent and Trademark Office (USPTO). For US-based inventors, we combine these patent data with demographic information from the US population censuses between 1856 and 1940. Finally, for US assignees, we add information on industrial research labs. How we link these datasets together is visualized in Fig. 6.

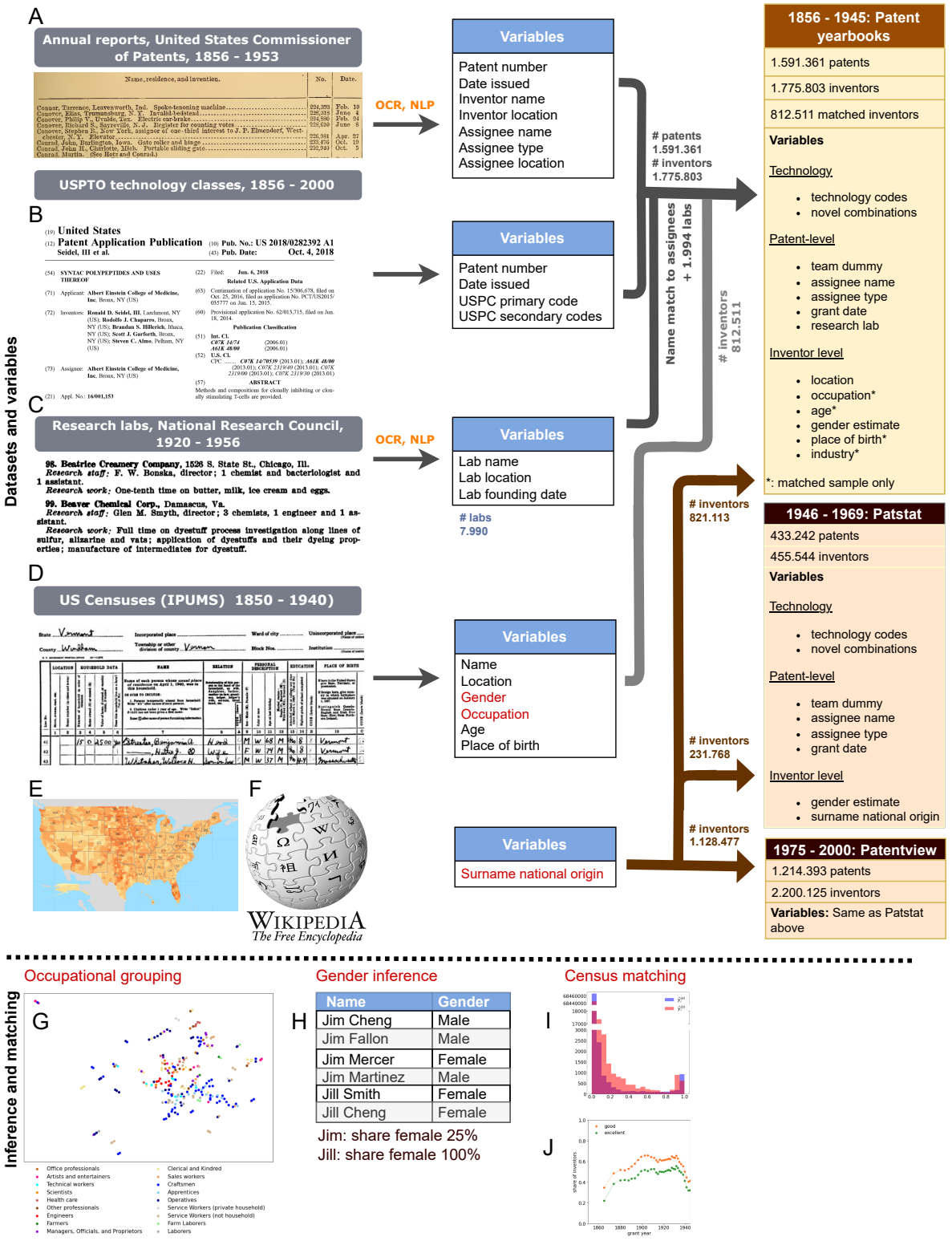


Figure 6: **A**: Sample page of the 1880 Annual report of the Commissioner of Patents. **B** First page of patent number 2018/0282392 A1. **C** Sample page of 1931 edition of "Industrial Research Laboratories of the United States". **D** Sample page of U.S. Census form. **E** Hispanic surnames are identified using models trained on the 2000 U.S. Census [24]. **F** East-Asian surnames are identified using models trained on Wikipedia [25]. **G**: Projection of cross-occupation labor flow network on 2-dimensional UMAP embedding. **H** First names are associated with a given gender based on a majority vote in over 600M census records. We focus on a sample where first names display little ambiguity in gender, requiring that at least 90% of census records agree on the gender. **I**: Histogram of $\hat{y}_i^{(a)}$ (blue) and $\hat{y}_i^{(b)}$ (red) in a random sample of about 68M potential matches. Some potential matches of seemingly high quality when using only name- and geographical distances are downgraded in the second-stage xgboost model, which also accounts for the quality of alternative matches. **J**: Match rate: share of inventors that can be matched with moderate accuracy ($\hat{y}_i^{(b)} \geq .95$, orange) and high accuracy ($\hat{y}_i^{(b)} \geq .99$, green) to one of the two census waves closest in time to the patent's grant date.

A.1.1 Patent yearbooks

We obtain patent data from three different sources. From 1856 until 1953 we use the *Annual Reports by the Commissioner of Patents*, henceforth referred to as *(patent) yearbooks*. These yearbooks contain, for a given year, information on each patent granted by the USPTO, including the name and location of residence of every inventor, the individuals or organizations to whom the patent was assigned, as well as the location thereof, a short description of the patent, and, from 1856 on, the grant number and date (see Fig. 6A for a sample page). Because we need the grant number to look up a patent's technology classes, our analysis starts in 1856.

Yearbooks have been scanned by different universities.[1] For most years, we can therefore access multiple digital copies of the same yearbook. All image files for these scanned copies were accessed through the *Hathi Trust Foundation*, except for scans obtained from the *Smithsonian Foundation*, which were obtained directly from this foundation. In total, this yields about half a million scanned pages.

To convert scans to text, we apply image preprocessing and optical character recognition (OCR). Next, we correct errors in the extracted text by cross-validating the output across different scans of the same yearbook, using a majority vote to decide on the correct string.

Next, we use natural language processing (NLP) to separate inventor names, locations, assignees, descriptions, patent numbers and dates in the extracted strings. To do so, we first manually created a ground truth that separates the aforementioned entities, which can be found in the online code appendices. In some reports, first names are listed in full only the first inventor. However, these yearbooks typically contain additional lines that provide full first and last names for the other inventors. These lines can be identified by the substring "(See <name of first inventor>)". Whenever possible, we supplement first-name information from these lines. However, in some cases and years only initials are provided for second and subsequent inventors.

Patents can be assigned (i.e., transfer intellectual property rights) to individuals or organizations other than the original inventors. If this happens before or at the time of the grant, the yearbooks contain this information, following the substring "assigned to". To distinguish between patents assigned to organizations (e.g., "assigned to Wright Metal Incorporated") and patents assigned to individuals (e.g. "assigned to Benjamin Reece and Neary Claflin"), we once again rely on NLP, manually creating a ground truth training set for named entity extraction. The vast majority of organizations in this period, over 99%, are firms.

Because the same organization may show up in different yearbooks with different names (e.g. "General Electrics" versus "General Electric Inc." versus "General Electric

[1]These are the following universities: Harvard University, Princeton University, University of California, University of Chicago, University of Illinois at Urbana-Champaign, University of Michigan, University of Wisconsin.

Incorporated") we manually align common terms (e.g., replace "Mfg." by "Manufacturing") and then apply a fuzzy name-matching algorithm that calculates the string similarity across all assignees. We use this to disambiguate organization names, merging names likely to refer to the same organizations.

This process allows us to extract detailed information on patents, their inventors and assignees for the period 1856-1953. In the analysis, we exclude the years 1873, 1874, 1878, 1908, 1909, 1951 and 1952. In 1874, yearbooks are missing altogether. In the other years, first names of inventors are not reported, which complicates the match to Census data. Furthermore, we restrict the sample of patents to utility patents, excluding reissues or translations of existing patents, provisional patents, design patents and (organic) plant patents. Finally, we drop patents where one or more inventors are located outside the US.

The Python code of these procedures - including ground truth training data sets - is available in the Supporting Material, Repository 1.

A.1.2 Recent patent data: EPO-PATSTAT and PatentsView

For the period 1954-2000, we use two datasets with pre-digitized patent and inventor information. For the period 1953-1969, we rely on the European Patent Office's EPO-PATSTAT database [26]. Inventor location data are missing in this period. For the period 1976-2000, we use the USPTO's own PatentsView database [27]. We exclude the period 1969-1975 from our analysis, because the EPO-PATSTAT data do not accurately report assignee and inventor information for these years: from 1969 on, individual assignees are no longer reliably reported and inventor countries of residence are missing altogether.

A.1.3 Census data

We obtain US Census records from the *Integrated Public Use Microdata Series*, or *IPUMS* [28]. These records, about 650 million in total, contain the answers to Census questions for all US residents for the years 1850, 1860, 1870, 1880, 1900, 1910, 1920, 1930 and 1940. 1890 is unavailable because a fire destroyed most of the records of that year. In our analysis, we use information on first and last names, years of birth, places of residence and occupations.

Matching inventors to the US census We match inventors to the US census between 1856 and 1945, using US Census records from the Integrated Public Use Microdata Series, or IPUMS [28]. To do so, we proceed in five steps:

1. For each inventor, we find a set of candidate matches based on the string distance between the inventor's last name and all last names in a US census wave. On average there are 326,697 candidate matches for each patent-inventor combination.

2. We create a ground truth for a subset of inventor-individual matches, relying on inventors whose patents are listed in Wikidata, along with detailed information on date and place of birth, places of residence, as well as spouses, children and parents.

3. We train a first-round xgboost model on this ground truth that predicts correct matches from information on distances between an inventor and all candidate matches in the census, using as predictors string distances for first names, last names and initials, as well as the geographical distance between the places of residence in IPUMS and on patents. This model provides for each candidate match-pair a score that describes the quality of each candidate match: $\hat{y}_a$.

4. We train a second-round xgboost model on a different set of ground truth observations that uses as predictors $\hat{y}_{ij}^{(a)}$, as well as the top k $\hat{y}_{ij'}^{(a)}$ scores across all match candidates associated with inventor i. This yields for each inventor-individual match a score $\hat{y}_{ij}^{(b)}$.

5. We link inventors to individuals by choosing the individual in the US census with the highest $\hat{y}_{ij}^{(b)}$ across all candidates j. The first- and second-round estimated quality of this match are denoted as $\hat{y}_i^{(a)}$ and $\hat{y}_i^{(b)}$ respectively.

The first-round model effectively decides how different distances between inventor and census individuals should be weighted when selecting between multiple candidate matches. The second-round model provides an estimate of the confidence in a match, taking into consideration that confidence should be low if either $\hat{y}_{ij}^{(a)}$ is low, or if there are multiple candidates j with more or less equal values of $\hat{y}_{ij}^{(a)}$.

Fig. 6I shows the histograms of $\hat{y}_i^{(a)}$ and $\hat{y}_i^{(b)}$ in a random sample of potential matches. In this paper, we only use matches for which $\hat{y}_i^{(b)} >= .99$. Although the second stage xgboost estimates, $\hat{y}_i^{(b)}$, do not improve match quality, they do downgrade a number of seemingly high-quality matches, where inventors with common first and last names found multiple close matches in the census data. Because the multiplicity of good matches actually complicates selecting the right match, the second stage xgboost estimates correctly lower the estimated quality of such matches.

We repeat this analysis, matching each inventor to the two nearest census waves, except for patents after 1940, where inventors are matched only to the 1940 wave. Next, we select among these two matches the wave with the highest $\hat{y}_i^{(b)}$, conditionally on the associated individual being at least 16 years old. In case one of the two matches is younger than 16 years, we select the other match, provided that $\hat{y}_i^{(b)} \geq .95$. If the resulting matched individual is younger than 16 years, we drop the inventor from the analysis. For the analyses we present in this paper, we only use matches where $\hat{y}_i^{(b)} \geq .99$. The match rate at both levels of accuracy over time is given in Fig. 6J.

A.1.4 Industrial research lab surveys

In 1920, the National Research Council started publishing the series *Industrial Research Laboratories of the United States*. This series contains the results from surveys on research activities by US firms. Inclusion in these surveys required that a firm had a dedicated "laboratory" with "separate and permanently established research staff and equipment", but excluding "firms that indicated they only occasionally carry out research, using teams temporarily recruited for the purpose or assembled from their operating staffs" [p. 2 29, see also [30]].

Separate issues were published in 1920, 1927, 1931, 1933, 1938, 1940, 1946, 1948, 1950 and 1956. Each book contains the name of the lab, a short description of its main activity, the lab's city or address, the (managing) directors and important researchers and, in some editions, the lab's major equipment and number of employees (Fig. 6C). Furthermore, the 1940 and 1946 editions also contain the founding year of each lab. In our analysis, we use these data to assess whether a firm owned a research lab and, if so, the lab's founding year.

Scanned versions of this series were obtained from the *Hathi Trust Foundation*. We use OCR to digitize the contents of these books and subsequently apply NLP to extract the name and location of each lab. We then name-match the labs in the books to organizations ("assignees") in the patent data. First, we clean the names by removing common words such as 'company' and 'limited'. Next, we calculate the Jaro-Winkler string similarity distance between every lab and patent assignee. Finally, we match records if the distance is higher than 0.95 or if the distance is higher than 0.90 and at least one of the words in the names matches perfectly.[2]

This procedure yields a total of 1,994 unique assignees for which we can establish that they operated an industrial research lab. Note that this number presents an undercount: not all firms that operated industrial research labs were included in the survey and for firms that were, we may not have been able to match their research labs to the assignee names on patents. To determine when a given assignee started operating a research lab, we use information on the founding years from the 1940 and 1946 editions. For labs that are not reported in these editions, we set the founding date to the year of the first survey that mentions these labs. Finally, when labs are mentioned in editions that predate the founding year, we override the founding years by the year of the earliest edition that mentions the lab.

The Python code and output of this exercise are available in the Supporting Material, Repository 2.

[2]Similar procedures using other similarity measures, e.g. using the Levenshtein distance, with different thresholds, yield similar results.

A.1.5 Technology classes

The USPTO labels all patents with technology codes from the United States Patent Classification (USPC classes). Each patent lists one and only one primary (PR) class. This primary class "[...] is indicative of the invention as a whole or the main inventive concept using the claims as a guide."[3] We use the PR code to divide patents into broad classes in some of our robustness analyses. Furthermore, patents often also list secondary (SR) classes. These secondary classes are meant to support patent examiners in exploring the state of the art when determining the merits of a patent application. We use the combination of PR and SR classes (without making a distinction between the two types of classes) to determine a patent's novelty.

The primary and secondary technology classes for each patent are made available by the USPTO through its *CASSIS Patents Assignments File* and *Bulk Data Products* repository.[4] We can match the patents in the yearbooks to these data using the patent number. Errors in our OCR may lead to ambiguous and/or imperfect matches. In these cases, we supplement the patent number information with the patent's grant date to improve the match.

The USPC classification system is hierarchically nested. At the highest level of aggregation, it contains classes marked by three numerical digits. In our datasets, we identify 474 unique classes. Each class is further subdivided into subclasses. Classes tend to distinguish among technologies, subclasses "delineate processes, structural features, and functional features of the subject matter encompassed within the scope of a class."[5] There are almost 150,000 unique subclasses in our datasets. We refer to these subclasses as "6-digit" codes, even when some of them have longer codes. Whenever the USPTO updates its classification system, it retroactively updates the patent classes on each patent. This ensures high comparability over time, because technology codes are harmonized across the entire period of analysis.

We further aggregate technology classes using the *USPTO Historical Patent Data Files* [31] which extend the *NBER Patent Citation File* [32]. These files group USPC classes and subclasses into coherent, more aggregated categories. At the highest level of aggregation, this classification distinguishes among six categories: *Mechanical*; *Chemical*; *Electrical & Electronic*; *Computers & Communication*; *Drugs & Medical*; and *Other*.

The NBER classification is not available for all patents in our data, and is more often missing in the most recent years. To remedy this, we recover the correspondence between the NBER subcategories and the USPC main classes empirically, using the primary technology classes for patents that list both NBER and USPC classes.

[3] `https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf`, p I.5.

[4] `https://bulkdata.uspto.gov/`

[5] `https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf`, p I.1.

A.2 Variable construction

A.2.1 Novelty

In the main text, we analyze to what extent different types of inventors and coordination mechanisms lead to inventions of differing novelty. To assess the novelty of a patented invention, scholars have adopted a variety of approaches, ranging from qualitative assessments to analyses of forward and backward citations [9]. The downside of these approaches is that scale poorly – as in the case of qualitative assessment – or that they require data that are not available for the entire period we analyze in this paper – as in the case of patent citations, which are only commonly used after the 1950s. Instead, we determine the novelty of a patented invention from the technology classes that they list.

Technology classes have the advantage that they are available in a harmonized classification system for the entire period of our study. To assess whether a patented invention is novel, we ask whether the exact combination of technology classes had ever been used before on a patent. This metric was originally proposed by [7] [see also, 8, 10]. Furthermore, we distinguish between radically novel combinations and incremental novelty by distinguishing between novel combinations at different levels of technological aggregation. In particular, if a patent lists 3-digit technology classes that had not been combined before, we view the patent's novelty as "radical". In contrast, if the combination technologies on the patent already existed at the level of 3-digit technology codes, but not at the level of 6-digit codes, we view the patent's novelty as "incremental".

A.2.2 Occupations

The census data contain harmonized occupation codes in the *occ1950* classification created by IPUMS from the original strings that the census enumerator provided. With over 250 different job titles, the occ1950 classification is too detailed to use in the descriptive analyses of this study. Therefore, we aggregate these job titles into broader classes. One possible concern is that the popularity of some occupational titles changed over time, even when jobs did not change substantially. In particular, occupations as reported in the census are ultimately based on self-reported strings. As a consequence, the emergence of engineers in the census may not only be due to an actual expansion of engineers in the population, but also because more individuals start describing themselves as engineers. To address this concern, we group occupations into classes, based of an analysis of labor flows between the detailed occupations of the occ1950 classification.

Flow-based grouping. To create labor-flow-based occupation groupings, we first create a matrix F with elements F_{ij} that contains the total number of individuals who move from from occupation i to occupation j in the subsequent census wave. To do so, we start by selecting all individuals with non-missing occupation codes. Next, some occupations were not always collected at the same level of detail. This is the case for *Professors* (codes

10-29) and *Scientists* (codes 61-69), which in some census waves are subdivided by field, which we aggregate into two higher level classes.

Next, we construct for each pair of sequential decades the total number of individuals who listed occupation i in the first decade and occupation j in the next decade. This provides us with decadal cross-occupational labor flows. Note that due to the loss of the 1890 census in a fire, labor flows that start in 1880 end in 1900. Because the population grows from 23M in 1850 to 132M in 1940, these counts will be dominated by job switches between later decades. To remedy this, we normalize the flows in each decade by the sum total of all flows in that decade to express them as shares that add up to 1 in each decade. Next, we multiply these shares with the grand total of all flows across all decades divided by eight, the number of decade pairs. This ensures that flows from each year are weighted equally, while retaining the total number of job switchers across all decades.

To turn the flow matrix F into a matrix of flow intensities, we calculate a quantity called *skill relatedness* [33]. Skill relatedness quantifies whether an observed flow between two occupations surpasses a random benchmark. Here, we follow [34], who develop an information-theoretic framework to generate Bayesian estimates of the amount of surprise involved in observing flows F_{ij}, given the total inflows into occupation j and the total outflows from occupation i. To be precise, we will estimate the point-wise mutual information $pmi(i,j) = \log \frac{q_{ij}}{q_i q_j}$, where q_{ij} is the probability of observing an individual move from occupation i to j, q_i the (marginal) probability that an individual leaves occupation i and q_j the marginal probability that an individual moves to occupation j. We collect these estimates in a matrix PMI.

Next, following recommendations of [35] and [36], we keep all elements of PMI that are significantly ($p = 0.01$) larger than 0, which we interpret as a sign of statistically significant skill relatedness between two occupations.[6] Next, we convert this measure of relatedness in a measure of distance by subtracting matrix PMI from the maximum value across all its elements. Finally, we set all elements in this distance matrix, D, that correspond to insignificant or negative relatedness to a value of ten times the maximum distance.

Using this distance matrix, we estimate a 10-dimensional *Uniform Manifold Approximation and Projection* [UMAP, 37] embedding on which we now project all occupations (see Fig. 6G for a projection on a 2-dimensional embedding). Finally, we use the Python implementation of Campello et al.'s [2013] *HDBSCAN* algorithm to cluster occupations at two different, nested hierarchical levels. The resulting clusters are provided in Table 1.

[6]Note that we retain the direction of the flows and do not symmetrize this relatedness matrix. Moreover, diagonal elements are treated the same as any other element.

Table 1: Occupational classification - flow-based grouping

Mailmen

Mailmen: 84: Express messengers and railway mail clerks; 85: Mail carriers; 91: Telegraph operators; 93: Ticket, station, and express agents

Transport Services

Transport Services: 64: Conductors, railroad; 78: Baggagemen, transportation; 83: Dispatchers and starters, vehicle; 179: Brakemen, railroad; 180: Bus drivers; 182: Conductors, bus and street railway; 195: Motormen, street, subway, and elevated railway; 204: Switchmen, railroad; 225: Guards, watchmen, and doorkeepers; 230: Policemen and detectives; 236: Watchmen (crossing) and bridge tenders

Printing

Printing: 106: Bookbinders; 112: Compositors and typesetters; 116: Electrotypers and stereotypers; 117: Engravers, except photoengravers; 147: Photoengravers and lithographers; 151: Pressmen and plate printers, printing; 172: Apprentices, printing trades

Logging

Logging: 29: Foresters and conservationists; 52: Surveyors; 124: Inspectors, scalers, and graders, log and lumber; 201: Sawyers; 246: Lumbermen, raftsmen, and woodchoppers

White Collar

Government: 67: Inspectors, public administration; 70: Officials and administrators (n.e.c.), public administration; 96: Auctioneers; 228: Marshals and constables; 233: Sheriffs and bailiffs

Financial: 1: Accountants and auditors; 79: Bank tellers; 80: Bookkeepers; 81: Cashiers; 87: Office machine operators; 94: Clerical and kindred workers (n.e.c.)

Other White Collar: 7: Authors; 10: Clergymen; 11: Professors; 17: Editors and reporters; 28: Farm and home management advisors; 32: Librarians; 44: Recreation and group workers; 45: Religious workers; 46: Social and welfare workers, except group; 48: Psychologists; 50: Miscellaneous social scientists; 53: Teachers (n.e.c.); 62: Buyers and department heads, store; 66: Floormen and floor managers, store; 71: Officials, lodge, society, union, etc.; 72: Postmasters; 74: Managers, officials, and proprietors (n.e.c.); 75: Agents (n.e.c.); 76: Attendants and assistants, library; 82: Collectors, bill and account; 89: Stenographers, typists, and secretaries; 95: Advertising agents and salesmen; 99: Insurance agents and brokers; 101: Real estate agents and brokers; 102: Stock and bond salesmen; 103: Salesmen and sales clerks (n.e.c.)

Continued on next page

Table 1: Occupational classification - flow-based grouping (Continued)

N.E.C.: 31: Lawyers and judges; 39: Personnel and labor relations workers; 47: Economists; 49: Statisticians and actuaries; 65: Credit men; 92: Telephone operators

Farming

Farming: 60: Farmers (owners and tenants); 61: Farm managers; 238: Farm foremen; 239: Farm laborers, wage workers; 240: Farm laborers, unpaid family workers

Engineers

Engineers: 4: Architects; 16: Draftsmen; 18: Engineers, aeronautical; 19: Engineers, chemical; 20: Engineers, civil; 21: Engineers, electrical; 22: Engineers, industrial; 23: Engineers, mechanical; 24: Engineers, metallurgical, metallurgists; 25: Engineers, mining; 26: Engineers (n.e.c.); 181: Chainmen, rodmen, and axmen, surveying

Service Work

Health Care: 9: Chiropractors; 13: Dentists; 37: Optometrists; 38: Osteopaths; 40: Pharmacists; 42: Physicians and surgeons; 54: Technicians, medical and dental; 55: Technicians, testing; 57: Therapists and healers (n.e.c.)

Low Skill Service: 15: Dietitians and nutritionists; 34: Nurses, professional; 35: Nurses, student professional; 77: Attendants, physicians and dentists office; 97: Demonstrators; 104: Bakers; 184: Dressmakers and seamstresses, except factory; 187: Fruit, nut, and vegetable graders, and packers, except factory; 190: Laundry and dry cleaning operatives; 192: Milliners; 210: Housekeepers, private household; 211: Laundresses, private household; 212: Private household workers (n.e.c.); 213: Attendants, hospital and other institution; 214: Attendants, professional and personal service (n.e.c.); 216: Barbers, beauticians, and manicurists; 217: Bartenders; 218: Bootblacks; 219: Boarding and lodging house keepers; 220: Charwomen and cleaners; 221: Cooks, except private household; 222: Counter and fountain workers; 223: Elevator operators; 226: Housekeepers and stewards, except private household; 227: Janitors and sextons; 229: Midwives; 231: Porters; 232: Practical nurses; 235: Waiters and waitresses; 237: Service workers, except private household (n.e.c.)

N.E.C.: 143: Opticians and lens grinders and polishers; 244: Gardeners, except farm, and groundskeepers

Arts And Design

Arts And Design: 5: Artists and art teachers; 14: Designers; 41: Photographers; 98: Hucksters and peddlers; 121: Furriers; 122: Glaziers; 158: Tailors and tailoresses; 198: Photographic process workers

Entertainment

Continued on next page

Table 1: Occupational classification - flow-based grouping (Continued)

Artists: 2: Actors and actresses; 12: Dancers and dancing teachers; 33: Musicians and music teachers

Sports And Farm Work: 6: Athletes; 27: Entertainers (n.e.c.); 51: Sports instructors and officials; 58: Veterinarians; 63: Buyers and shippers, farm products; 191: Meat cutters, except slaughter and packing house; 241: Farm service laborers, self-employed

N.E.C.: 215: Attendants, recreation and amusement

Shipyard Work

Nautical: 69: Officers, pilots, pursers and engineers, ship; 178: Boatmen, canalmen, and lock keepers; 200: Sailors and deck hands; 242: Fishermen and oystermen

Laborers: 245: Longshoremen and stevedores; 248: Laborers (n.e.c.)

N.E.C.: 111: Cement and concrete finishers

Textile Workers

Textile Workers: 131: Loom fixers; 185: Dyers; 202: Spinners, textile; 207: Weavers, textile; 209: Operative and kindred workers (n.e.c.)

Blue Collar

Locomotive Workers: 118: Excavating, grading, and road machinery operators; 129: Locomotive engineers; 130: Locomotive firemen; 155: Stationary engineers; 203: Stationary firemen; 224: Firemen, fire protection

Oilers And Hoistmen: 113: Cranemen, derrickmen, and hoistmen; 196: Oilers and greaser, except auto

Mining And Railroads: 125: Inspectors (n.e.c.); 137: Mechanics and repairmen, railroad and car shop; 177: Blasters and powdermen; 193: Mine operatives and laborers; 194: Motormen, mine, factory, logging camp, etc.

Carpentry: 110: Carpenters; 140: Millwrights; 146: Pattern and model makers, except paper; 165: Apprentice carpenters

Metal Workers: 105: Blacksmiths; 120: Forgemen and hammermen; 123: Heat treaters, annealers, temperers; 141: Molders, metal; 152: Rollers and roll hands, metal; 188: Furnacemen, smeltermen and pourers; 189: Heaters, metal

Machine Workers: 127: Job setters, metal; 132: Machinists; 160: Tool makers, and die makers and setters; 167: Apprentice machinists and toolmakers

Continued on next page

Table 1: Occupational classification - flow-based grouping (Continued)

	Repair Office Machinery: 43: Radio operators; 135: Mechanics and repairmen, office machine; 136: Mechanics and repairmen, radio and television
	Electrical: 115: Electricians; 128: Linemen and servicemen, telegraph, telephone, and power; 142: Motion picture projectionists
	Motorized Transportation: 3: Airplane pilots and navigators; 133: Mechanics and repairmen, airplane; 134: Mechanics and repairmen, automobile; 163: Apprentice auto mechanics
	Motorized Transportation: 183: Deliverymen and routemen; 205: Taxicab drivers and chauffers; 206: Truck and tractor drivers; 243: Garage laborers and car washers and greasers; 247: Teamsters
	Masonry: 108: Brickmasons, stonemasons, and tile setters; 149: Plasterers; 156: Stone cutters and stone carvers; 164: Apprentice bricklayers and masons
	Other Construction: 144: Painters, construction and maintenance; 145: Paperhangers; 150: Plumbers and pipe fitters; 159: Tinsmiths, coppersmiths, and sheet metal workers; 161: Upholsterers; 170: Apprentices, building trades (n.e.c.); 173: Apprentices, other specified trades; 174: Apprentices, trade not specified; 197: Painters, except construction or maintenance
	N.E.C.: 90: Telegraph messengers; 100: Newsboys; 107: Boilermakers; 109: Cabinetmakers; 114: Decorators and window dressers; 138: Mechanics and repairmen (n.e.c.); 148: Piano and organ tuners and repairmen; 154: Shoemakers and repairers, except factory; 162: Craftsmen and kindred workers (n.e.c.); 166: Apprentice electricians; 168: Apprentice mechanics, except auto; 169: Apprentice plumbers and pipe fitters; 171: Apprentices, metalworking trades (n.e.c.); 175: Asbestos and insulation workers; 176: Attendants, auto service and parking; 186: Filers, grinders, and polishers, metal; 199: Power station operators; 208: Welders and flame cutters; 234: Ushers, recreation and amusement
N.E.C.	
	N.E.C.: 8: Chemists; 30: Funeral directors and embalmers; 36: Scientists; 56: Technicians (n.e.c.); 59: Professional, technical and kindred workers (n.e.c.); 68: Managers and superintendents, building; 73: Purchasing agents and buyers (n.e.c.); 86: Messengers and office boys; 88: Shipping and receiving clerks; 119: Foremen (n.e.c.); 126: Jewelers, watchmakers, goldsmiths, and silversmiths; 139: Millers, grain, flour, feed, etc.; 153: Roofers and slaters; 157: Structural metal workers

The flow-based clustering of occupations yields a group of occupations that contains all engineering occupations. However, this group also includes three further occupations that have strong labor-flow connections to engineering jobs: *Architects*, *Draftsmen* and *Chainmen, rodmen, and axmen, surveying* (i.e., occupations related to land surveying).

Table 2: Occupational classification - hierarchical grouping

Codes	**Class name**
0 - 99	Professional, technical
41 - 49	Engineers
7, 10, 12-19, 23-29, 61-69, 81-84	Scientists
1, 4, 6, 31, 51, 57, 74, 77, 91	Artists
8, 32, 34, 58-59, 70-71, 73, 75, 97-98	Health care
0, 9, 55, 72	Office professionals
2-3, 33, 92, 94-96	Technical workers
36, 52-54, 56, 76, 78-79, 93, 99	Other professionals
100 - 123	Farmers
200 - 290	Managers, Officials, and Proprietors
300 - 390	Clerical and Kindred
400 - 490	Sales workers
500 - 594	Craftsmen
600 - 690	Operatives
600 - 621	Apprentices
622 - 690	Operatives
700 - 720	Service Workers (private household)
730 - 790	Service Workers (not household)
810 - 840	Farm Laborers
910 - 970	Laborers
595, 979 - 999	Missing

Broad occupational sectors. To check the robustness of our results, we also consider a different way of grouping of occupations into broad sectors, using the hierarchical structure of the occ1950 codes. Because of our interest in the rise of engineers among inventors, we subdivide the sector "Professional, technical" into a number of smaller subclasses. Furthermore, we separate apprentices as a subclass of the sector "Operatives". This yields the following set of occupation sectors:

The missing category consists of the following reported occupation codes: *Members of the armed services*; *Not yet classified*; *Keeps house/housekeeping at home/housewife*; *Imputed keeping house* (1856-1900); *Helping at home/helps parents/housework*; *At school/student*; *Retired*; *Unemployed/without occupation*; *Invalid/disabled w/ no occupation reported*; *Inmate*; *New Worker*; *Gentleman/lady/at leisure*; *Other non-occupational response*; *Occupation missing/unknown*; and *N/A (blank)*.

We can now determine which occupations are overrepresented in the patent records,

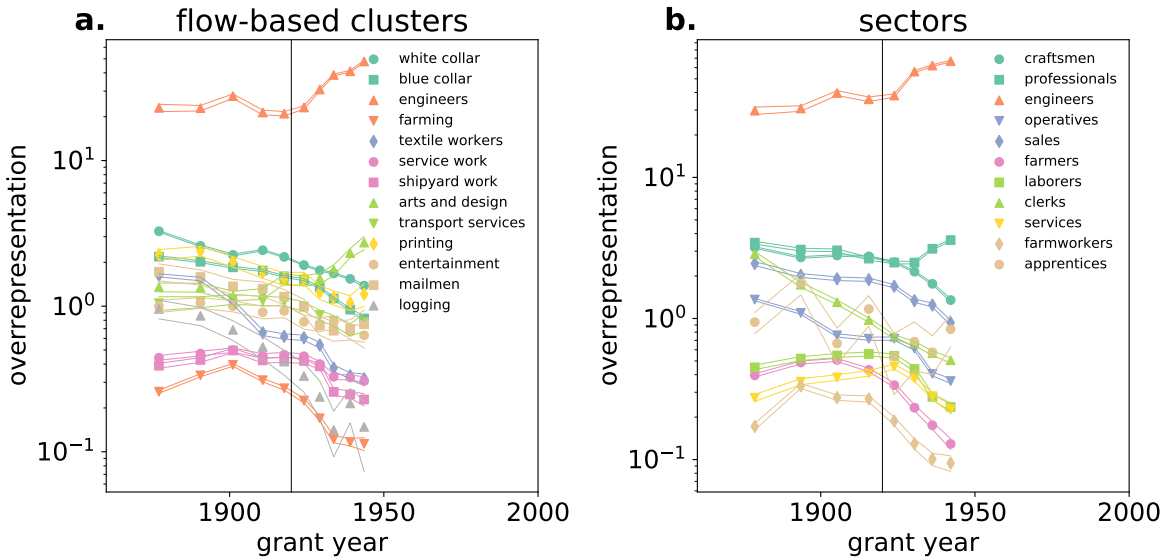


Figure 7: Overrepresentation of broad occupation classes in patent records. **a**: Overrepresentation of flow-based occupational clusters. **b**: Overrepresentation of occupational sectors.

by comparing their shares among inventors to their shares in the working-age population. To do so, we calculate the following quantity:

$$\rho_{ot} = \frac{N_{ot}/\sum_{o'} N_{o't}}{POP_{ot}/\sum_{o''} POP_{o''t}} \tag{1}$$

where N_{ot} is the number of patents by inventors with occupation o in period t and POP_{ot} the number of workers in the census records with occupation o in period t. To determine POP_{ot} outside census years, we interpolate linearly between two census waves. Fig. 7 plots the overrepresentation of flow-based clusters (7a) and of broad occupational sectors (7b) over time.

The Python code of this exercise is available in the Supporting Material, Repository 3.

A.2.3 Distinguishing family teams

There are two ways in which we can assess in our dataset whether co-inventors are related. First, for patents until 1953, the matched census records allow us to construct family ties. This approach has the advantage that it allows identifying a large set of family ties of different degrees (e.g., uncles and nieces). However, the construction of these family ties relies on matching inventors to the census, and for more distant family ties, linking census

individuals across multiple census waves. Using census linkages to determine family ties has the disadvantage that both of these types of linkages are imperfect and only available until 1945.

Instead, therefore, we rely on a second approach and identify family ties based on shared last names. That is, we assume that two inventors are related if they share the same last name. Because not all family members will share the same last name, this approach results in an under-count of family-based teamwork. Moreover, even when two inventors share the same last name, they are not necessarily related. Such spurious family ties will be particularly common when inventors have common last names. The nature of this problem changes over time. In the 1940s, the most common last name is *Smith* and inventors with that last name hold 0.78% of all patents. This is followed by *Johnson* (0.51%), *Miller* (0.47%), *Brown* (0.37%) and *Anderson* (0.35%). From the 1980s on inventors with last names that are common in certain East-Asian countries start dominating this list. For instance, the top 5 surnames on patents in 2010 is composed of *Kim* (1.17%), *Lee* (1.15%), *Chen* (.86%), *Wang* (0.77%) and *Park* (0.61%). This reflects the fact that in some countries (especially in Asia), a small number of surnames is very frequent. For instance, in our data, we find that 19.9% of inventors residing in Korea record the surname *Kim* and the five most common Korean surnames account for 49.3% of all inventors in Korea. Similarly high percentages are found for inventors residing in Taiwan (top 5 surnames: 33.1% of total), China (30.5%), Malaysia (19.7%), Hong Kong (18.8%) and Denmark (14.4%).

To correct our estimates of family-based team-patents, we create a random benchmark in which we shuffle inventor surnames across patents. However, we do this in such a way that inventors whose name suggests a certain "ethnicity" can only swap names inventors with the same inferred ethnicity.

To construct this ethnicity variable, we drop all inventors residing in the US, Canada or Australia, because the populations of these countries have always included many migrants from all over the world. Furthermore, because China, Taiwan, Hong Kong, Singapore and Malaysia share many of the same last names, we group the inventors residing in these countries into a single category. Similarly, we group all inventors in Spain and Spanish-speaking Latin American countries, as well as Germany, Austria and Liechtenstein. Finally, countries with fewer than 100,000 patents are grouped in a residual category, *RoW*. For every given surname, we now determine the most likely geographical origin as the country that accounts for the greatest share of inventors with that surname.[7]

Finally, we randomize surnames across inventor-patent combinations. To do so, we shuffle the surname column within (grant-year, name-origin) pairs. Fig. 8 shows that for the period reported in Fig. 3, the null-model shares are negligible compared to those of observed family-based teams in- and outside firms.

[7]This procedure leads to some problems for Indian surnames, which are also very common in the UK. For these names, we make a number of manual adjustments that can be found in the code in the accompanying Supporting Material.

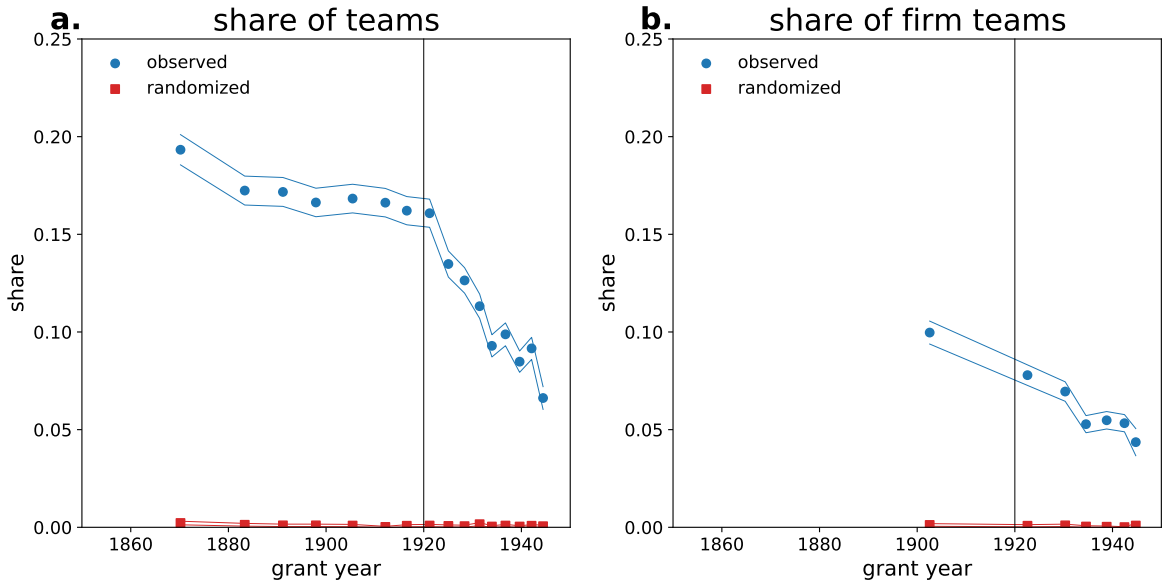


Figure 8: Null model comparison family teams. **a**: Share of teams that list inventors with the same last name. **b**: Share of firm patents that list inventors with the same last name. Blue: shares in observed data, red: shares in reshuffled data. Thin lines display 95% confidence intervals.

The Python code of this exercise is available in the Supporting Material, Repository 4.

A.2.4 Disambiguated inventor dyads

Disambiguating inventor names is complicated, given that some last names are so common that they represent a significant share of the population. Adding to this that the same person may report slightly different names on a patent than in the census (a problem compounded by spelling mistakes and OCR errors), we consider disambiguating inventors across patents as beyond the scope of this paper. In contrast, however, disambiguating pairs of co-inventors is relatively easy. To see this, let p_n denote the share of people in the US with name n. The probability of observing a patent by a random pair of inventors a and b is therefore $p_a * p_b$, a number that is typically very small. Therefore, if we observe inventor-pair (a, b) more than once, it is very likely that this marks a repeated collaboration.

To assess the share of pairs of inventors that enter in repeated collaborations, we define dyads as combinations of two last names. We focus on inventors whose last name is repeated fewer than 500k times across all 9 US censuses. Because there are roughly 650M individuals in theses censuses, if we pick two random individuals, the probability of drawing any given combination of two last names is at most $p = \left(\frac{1}{1{,}300}\right)^2 = \frac{1}{1.69M}$.

Furthermore, we drop all inventor pairs who share the same last name if this last name is repeated over 50k times across all 9 US censuses, because family ties lead to correlated draws. Finally, we drop inventors with East-Asian or Indian surnames. This yields a total of 133k unique inventor dyads across all patents from 1856 to 1949. For 99.9% of repeated dyads, the time between first and last patent is shorter than 24 years, suggesting that most dyads are plausibly made up of the same two inventors.

In Fig. 3c, we analyze the stability of inventor dyads. To do so, we focus on the set of patents that list inventor dyads that could be disambiguated, following the steps outlined above. We call this set S_{disamb}. Next, for each decade, we calculate how many patents within this set resulted from repeated collaborations within that decade. That is, we mark all dyads that patent more than once in a given decade and then calculate how many patents in that same decade belong to such dyads. Next, we divide this by the total number of patents in S_{disamb} granted in the decade. Fig. 3c repeats this process for patents assigned to different types of assignees.

A.2.5 Gender

To assess the gender of inventors, we could rely on the matched census records. This approach has two disadvantages. First, we would only be able to determine a gender for the inventors that we can accurately match to the census records. Second, and more importantly, our census records only allow us to match inventors until 1945. Instead,

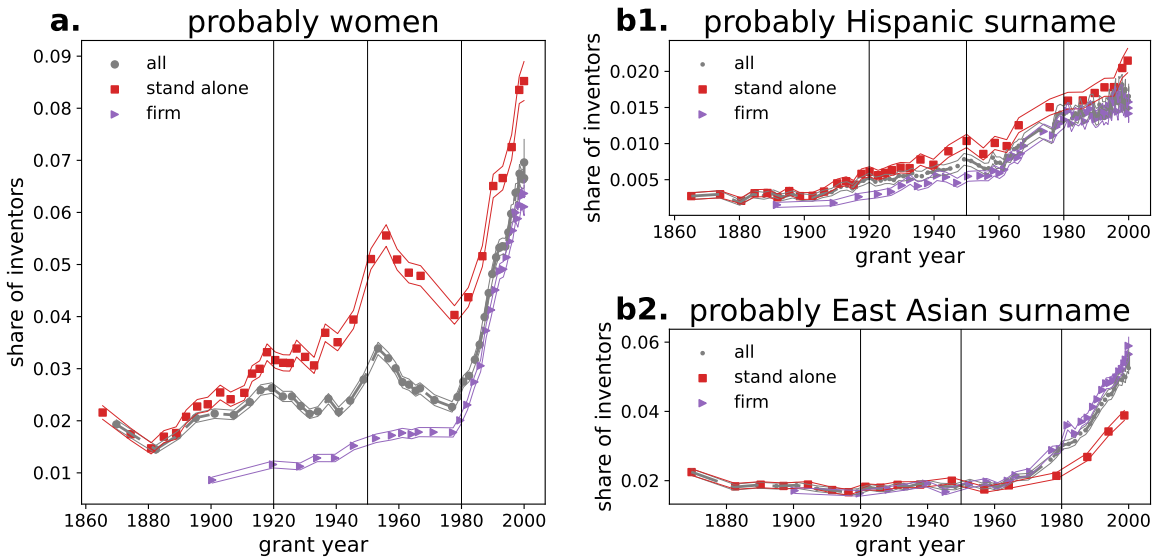


Figure 9: Robustness participation analysis. **a**: Share of inventors with a first name that is used in at least 50% of census records by individuals that list their gender as female. **b**: Share of surnames that are probably of Hispanic origin (i.e., with positive but below 90% probability). **c**: Share of surnames that are probably of East-Asian origin (i.e., with positive but below 90% probability). Gray: all patents, red: stand-alone patents, purple: firm patents.

therefore, we infer the most likely gender of an inventor based on their first names. To do so, we calculate which share of individuals in the US censuses between 1850 and 1940 with a given first name listed their gender as female or male. Whenever this percentage in the US population (not among inventors) exceeds 90%, we infer that the first name is mostly used by women, respectively men. In these cases, we label inventors with such first names by the most common gender for these names in the census records. Furthermore, we label gender as missing if the first name is shorter than 3 characters, to avoid inferring gender from initials that were misidentified by our entity recognition.

For first names that are often used by both men and women (i.e., where less than 90% of individuals list the same gender), we do not infer a gender and instead set this variable to missing. To assess how important this choice is, we also analyze the share of inventors with first names that are associated with female genders in the census in 50% or more of all cases (Fig 9a). This analysis corroborates the results in Fig 5.

A.2.6 Surname national origins

To infer the national origins of surnames, we rely on predictions[8] from models that have been trained on datasets that contain surnames and their associated nationality by [24]. We focus on Hispanic and East-Asian surnames, because these are relatively easy to identify. Hispanic surnames are identified from models trained on data from the U.S. Census held in 2000, and East-Asian surnames are identified from models trained on Wikipedia data. [25].

Both algorithms yield an accuracy score that provides an estimated probability that a given surname indeed is Hispanic or East-Asian respectively. Whenever this score doesn't surpass 0.9, we refrain from labeling the surname as Hispanic or East-Asian. As a consequence, the estimated inventor shares are undercounts. However, our main interest is in how these shares differ between firm-based and stand-alone patents and how this difference evolves over time. For this purpose, we only need to assume that the degree of undercounting does not vary in a systematic way over time or with whether or not an inventor patents on behalf of a firm.

To assess how robust this analysis is, we repeat the analysis for individuals with surnames that are *probably* of Hispanic or East-Asian origin (Fig 9b and 9c). That is, we calculate the shares of inventors whose surname suggest one of these two origins with a probability that is nonzero, yet below 90%. The findings are in line with those reported in Fig. 5.

A.3 Graphs

A.3.1 General

The graphs in the main text typically show how quantities change over time. Because the number of patented inventions per year increases roughly exponentially, we face the problem of choosing proper time windows over which to average observations such that we balance a sufficiently high temporal resolution with reasonably precise point estimates.

To do so, we create windows, not based on time, but on temporal rank. That is, we sort all patents by their exact grant date and then divide the data into groups of N observations. Next, we calculate the average grant year associated with each group and use that as the horizontal coordinate in the graph. Meanwhile, the averages for the variable of interest, along with a 95% confidence interval are plotted along the vertical axis. This confidence interval is calculated as $\pm 1.96 \times$ `(standard error of the mean)`. Note that the group size differs across, and sometimes even within graphs. This is due to the fact that some types of observations are more numerous than others. For instance, patents by research labs are rather rare, especially in earlier years, whereas firm-based patents are found relatively often across most decades. Therefore, striking a useful balance

[8] Obtained using the *ethnicolr* Python package by [24]

between precision of point estimates and time resolution leads to different group sizes for these two types of patents.

The exception to this approach is Fig. 3c, where data are collected within a decade, Fig. 3,d1 and d2, where data are collected by period, and Fig. 4 where we use 11-year moving time-windows.

A.3.2 Geography graph

Fig. 4 corrects panels a and c for the distribution of population across cities. In panel a, we do this by dividing the effective number of cities for the patent distribution by the effective number of cities for distribution of the US population across cities. In panel c, we correct for population dynamics by first creating locational vectors. In particular, in each time window, we limit the set of cities to those that account for at least 0.1% of the overall US population. Next, we divide the share of patents a city holds by the share of population it hosts. This locational vector captures the overrepresentation of a city in inventive activity:

$$\rho_{cst} = \frac{N_{cst} / \sum_{c'} N_{c'st}}{POP_{ct} / \sum_{c''} POP_{c''t}} \tag{2}$$

where N_{cst} is the number of patents in city c, system s and period t, and POP_{ct} the population in city c and period t. To determine POP_{ct} outside census years, we interpolate linearly between two census waves.

We repeat this procedure once for the patents of system 1 and once for the patents of system 2. Next, we calculate the correlation between the locational vectors of system 1 and 2, which we then plot in Fig. 4c.

A.4 Additional results

This section presents a number of additional results in support of the analysis in the main text.

A.4.1 Regression analysis

In this section, we analyze whether certain inputs and coordination mechanisms are more likely to be used when a patent lists a novel combination of technologies. That is, the outcome variable is a dummy variable that takes on a value of 1 if the combination of technologies on the patent had not been used before and 0 if one or more prior patents with this exact same combination of technologies had already existed:

$$y_p = \begin{cases} 1, & \text{if combination of technologies is new} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

All variables of interest in these analyses are dummies or dummy groups. Our main specifications consist of models that are fully saturated or almost saturated in these covariates. As a consequence, the conditional expectation function (CEF) is linear and the best unbiased estimate is provided by Ordinary Least Squares analysis [39]). Therefore, we rely on Linear Probability Models (LPMs). However, in robustness checks, we control for year fixed effects. Because these models are not fully saturated, the CEF is not linear in the regressors. However, under these circumstances, our LPMs still provide the best linear approximation of the CEF.

In the regression models, we try to attribute the likelihood that a patent introduces a new technological combination to the labor inputs (engineers and teams) and coordination modes (firms and industrial research labs) of system 2. To proxy labor inputs, we ask whether or not at least one inventor on the patent is an engineer. Furthermore, we ask whether the patent was produced by a solo inventor or by a team of inventors. To assess how inventive activity was coordinated, we distinguish between patents that were assigned to firms, patents that were assigned to firms with research labs and patents that were assigned to individuals (either the inventors themselves or other individuals). Our prime goal is to assess whether certain types of organization enhance the capacity of innovation inputs to produce novel combinations. That is, we are interested in interaction effects between inputs and organizational arrangements.

Furthermore, we distinguish between effects on radical and incremental novelty. To do so, we run all analyses twice, once using novelty at the level of 3-digit technology codes and once at the level of 6-digit technology codes.

Finally, to determine whether estimated parameters change over time, we split our data into four periods: the period before the take-off of research labs (1856-1919), the period in which research labs start dominating invention (1920-1945) and the periods 1946-1969 and 1976-2000. Results are summarized in Tables 3-9.

To assess the robustness of these findings, we also run models that add year fixed effects and year-NBER-sector fixed effects, where the latter interact year dummies with the 6 technological sector dummies in [40]. Furthermore, so far, we have grouped all assignees other than individuals into one category, which we labelled firms. However, some of these assignees are actually better classified as other types of organizations, such as universities and government agencies. This is only an issue after 1945: before 1945, fewer than 1% of organizational patents are assigned to other organizations than firms, such as universities and government agencies. To test the robustness of our results, we drop all patents that were assigned to non-firm organizations and rerun our regression analyses. This yields the following alternative specifications:

(A) baseline model (Table 3):

1. dummy for whether or not one of the inventors is an engineer;
2. dummy for whether or not the patent lists a team of inventors;

3. interaction of 1 and 2;
4. dummy for whether or not the patent was assigned to a firm;
5. dummy for whether or not the patent was assigned to a firm with a research lab;
6. interaction of 1 and 4;
7. interaction of 1 and 5;
8. interaction of 1, 2 and 4;
9. interaction of 1 and 5;
10. interaction of 2 and 5;
11. interaction of 1, 2 and 5;

(B) baseline model + year fixed effects (Table 4);

(C) baseline model + year × technological sector fixed effects (Table 5);[9]

Next, we analyze a smaller model to study how novelty effects change over the entire time period from 1856 to 2000. Because we don't observe demographic information and research labs in this period, we limit the analysis to the effects of teams and firms on the likelihood that a patent lists a novel combination of technology codes.

(A*) baseline model (Table 6):

1. dummy for whether or not the patent lists a team of inventors;
2. dummy for whether or not the patent was assigned to a firm;
3. interaction of 1 and 2;

(B*) baseline model + year fixed effects (Table 7);

(C*) baseline model + year × technological sector fixed effects (Table 8);[10]

(D*) baseline model after dropping patents assigned to non-firm organizations (Table 9).

The main findings in the regression analyses can be summarized as follows:

- Engineers
  - Engineers tend to generate more novel combinations than non-engineers
  - This finding holds across periods and in models with year or technology fixed effects

[9]Technological sector refers to the 6 high-level groupings in [32].
[10]Technological sector refers to the 6 high-level groupings in [32].

Table 3: Novelty regression - Model A (baseline)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.023***		0.022***	0.023***	0.023***	0.023***	0.013***		0.015***	0.019***	0.019***	0.016***	0.018***		0.016***	0.022***	0.022***	0.023***
	(0.001)		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.003)	(0.003)	(0.001)		(0.001)	(0.003)	(0.003)	(0.003)
team		0.009***	0.008***	0.002**	0.002**	0.002**		0.001	0.001	0.001	0.001	0.001		0.013***	0.012***	0.005***	0.005***	0.005***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
engXteam			0.002	-0.000	-0.000				-0.014*	-0.025***	-0.025***				0.001	0.010	0.010	
			(0.003)	(0.006)	(0.006)				(0.006)	(0.007)	(0.007)				(0.003)	(0.008)	(0.008)	
firm				0.017***	0.010***	0.010***				0.002**	0.001	0.001				0.018***	0.011***	0.011***
				(0.000)	(0.000)	(0.000)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
engXfirm				-0.011***	-0.012***	-0.012***				-0.014**	-0.013*	-0.009				-0.014***	-0.014***	-0.017***
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.003)	(0.004)	(0.003)
teamXfirm				0.012***	0.010***	0.010***				0.003	0.003	0.003				0.009***	0.008***	0.008***
				(0.001)	(0.002)	(0.001)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)
engXteamXfirm				-0.001	-0.001					0.027*	0.028*					-0.012	-0.013	
				(0.007)	(0.007)					(0.013)	(0.013)					(0.009)	(0.009)	
RnDlab					0.031***	0.031***					0.019***	0.019***					0.026***	0.026***
					(0.001)	(0.001)					(0.003)	(0.003)					(0.001)	(0.001)
engXRnDlab					-0.012***	-0.011***					-0.014	-0.016					-0.009**	-0.008*
					(0.003)	(0.003)					(0.013)	(0.012)					(0.003)	(0.003)
teamXRnDlab					-0.000	0.000					0.004	0.003					-0.002	-0.001
					(0.003)	(0.003)					(0.011)	(0.010)					(0.003)	(0.003)
engXteamXRnDlab					0.004						-0.011						0.008	
					(0.008)						(0.033)						(0.008)	
Constant	0.054***	0.053***	0.053***	0.048***	0.048***	0.048***	0.044***	0.045***	0.044***	0.044***	0.044***	0.044***	0.064***	0.063***	0.062***	0.054***	0.054***	0.054***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.107***		0.105***	0.106***	0.106***	0.104***	0.064***		0.065***	0.084***	0.084***	0.079***	0.073***		0.073***	0.090***	0.090***	0.088***
	(0.002)		(0.002)	(0.004)	(0.004)	(0.004)	(0.005)		(0.005)	(0.007)	(0.007)	(0.006)	(0.002)		(0.003)	(0.005)	(0.005)	(0.005)
team		0.035***	0.032***	0.010***	0.010***	0.010***		0.005*	0.005*	0.002	0.002	0.002		0.048***	0.046***	0.026***	0.026***	0.025***
		(0.001)	(0.001)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
engXteam			-0.003	-0.011	-0.011				-0.011	-0.035	-0.035				-0.018**	-0.015	-0.015	
			(0.005)	(0.011)	(0.011)				(0.014)	(0.018)	(0.018)				(0.006)	(0.013)	(0.013)	
firm				0.078***	0.052***	0.052***				0.020***	0.015***	0.014***				0.069***	0.043***	0.043***
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.001)	(0.001)	(0.001)
engXfirm				-0.045***	-0.049***	-0.050***				-0.060***	-0.065***	-0.058***				-0.052***	-0.051***	-0.053***
				(0.005)	(0.005)	(0.005)				(0.011)	(0.011)	(0.010)				(0.006)	(0.007)	(0.006)
teamXfirm				0.041***	0.038***	0.037***				0.014**	0.017***	0.018***				0.020***	0.021***	0.019***
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.004)	(0.004)	(0.004)
engXteamXfirm				-0.003	-0.001					0.052	0.054					0.000	-0.004	
				(0.012)	(0.013)					(0.029)	(0.030)					(0.015)	(0.016)	
RnDlab					0.117***	0.117***					0.083***	0.083***					0.090***	0.090***
					(0.002)	(0.002)					(0.006)	(0.006)					(0.002)	(0.002)
engXRnDlab					-0.042***	-0.039***					0.018	0.015					-0.032***	-0.027***
					(0.006)	(0.005)					(0.027)	(0.025)					(0.006)	(0.005)
teamXRnDlab					-0.014**	-0.012**					-0.024	-0.025					-0.017**	-0.013**
					(0.005)	(0.004)					(0.021)	(0.020)					(0.005)	(0.005)
engXteamXRnDlab					0.015						-0.024						0.027*	
					(0.013)						(0.070)						(0.013)	
Constant	0.525***	0.525***	0.522***	0.499***	0.499***	0.499***	0.476***	0.476***	0.475***	0.472***	0.472***	0.472***	0.581***	0.580***	0.575***	0.545***	0.545***	0.545***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 4: Novelty regression - Model B (year fixed effects)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.010***		0.010***	0.017***	0.017***	0.017***	0.015***		0.016***	0.021***	0.021***	0.018***	0.009***		0.008***	0.015***	0.015***	0.017***
	(0.001)		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.003)	(0.003)	(0.001)		(0.001)	(0.003)	(0.003)	(0.003)
team		0.005***	0.005***	0.002*	0.002*	0.002*		0.001	0.001	0.001	0.001	0.000		0.008***	0.008***	0.003*	0.003*	0.004**
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
engXteam			0.002	-0.001	-0.001				-0.014*	-0.025***	-0.025***				0.003	0.011	0.011	
			(0.003)	(0.006)	(0.006)				(0.006)	(0.007)	(0.007)				(0.003)	(0.008)	(0.008)	
firm				0.010***	0.006***	0.006***				0.005***	0.004***	0.004***				0.013***	0.007***	0.007***
				(0.000)	(0.000)	(0.000)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
engXfirm				-0.014***	-0.014***	-0.014***				-0.015**	-0.014**	-0.010*				-0.013***	-0.014***	-0.016***
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.003)	(0.004)	(0.003)
teamXfirm				0.006***	0.006***	0.006***				0.003	0.003	0.003				0.005**	0.006**	0.005*
				(0.001)	(0.002)	(0.001)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)
engXteamXfirm				0.003	0.001					0.027*	0.027*					-0.010	-0.012	
				(0.007)	(0.007)					(0.013)	(0.013)					(0.009)	(0.009)	
RnDlab					0.021***	0.020***					0.020***	0.020***					0.020***	0.020***
					(0.001)	(0.001)					(0.003)	(0.003)					(0.001)	(0.001)
engXRnDlab					-0.007*	-0.006*					-0.015	-0.016					-0.006	-0.004
					(0.003)	(0.003)					(0.013)	(0.012)					(0.003)	(0.003)
teamXRnDlab					-0.002	-0.001					0.004	0.003					-0.003	-0.002
					(0.003)	(0.003)					(0.011)	(0.010)					(0.003)	(0.003)
engXteamXRnDlab					0.007						-0.011						0.008	
					(0.008)						(0.033)						(0.008)	
Constant	0.054***	0.054***	0.054***	0.051***	0.050***	0.050***	0.044***	0.045***	0.044***	0.044***	0.044***	0.044***	0.064***	0.064***	0.063***	0.058***	0.058***	0.058***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.046***		0.045***	0.070***	0.072***	0.069***	0.063***		0.065***	0.086***	0.086***	0.081***	0.041***		0.041***	0.064***	0.065***	0.063***
	(0.002)		(0.002)	(0.004)	(0.004)	(0.004)	(0.005)		(0.005)	(0.007)	(0.007)	(0.006)	(0.002)		(0.003)	(0.005)	(0.005)	(0.005)
team		0.019***	0.018***	0.009***	0.010***	0.009***		0.006**	0.006**	0.003	0.003	0.003		0.031***	0.030***	0.021***	0.021***	0.020***
		(0.001)	(0.001)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
engXteam			-0.004	-0.017	-0.017				-0.012	-0.036*	-0.036*				-0.010	-0.013	-0.013	
			(0.005)	(0.010)	(0.010)				(0.014)	(0.018)	(0.018)				(0.006)	(0.013)	(0.013)	
firm				0.034***	0.021***	0.021***				0.017***	0.012***	0.012***				0.046***	0.028***	0.028***
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.001)	(0.001)	(0.001)
engXfirm				-0.048***	-0.053***	-0.051***				-0.063***	-0.069***	-0.061***				-0.046***	-0.048***	-0.048***
				(0.005)	(0.005)	(0.005)				(0.011)	(0.011)	(0.010)				(0.006)	(0.006)	(0.006)
teamXfirm				0.018**	0.020***	0.019***				0.013*	0.016**	0.017**				0.008*	0.011**	0.010**
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.004)	(0.004)	(0.004)
engXteamXfirm				0.014	0.011					0.054	0.057					0.008	0.001	
				(0.012)	(0.013)					(0.029)	(0.030)					(0.014)	(0.016)	
RnDlab					0.071***	0.070***					0.075***	0.075***					0.067***	0.067***
					(0.002)	(0.002)					(0.006)	(0.006)					(0.002)	(0.002)
engXRnDlab					-0.020***	-0.016**					0.023	0.019					-0.020**	-0.014**
					(0.006)	(0.005)					(0.027)	(0.025)					(0.006)	(0.005)
teamXRnDlab					-0.016***	-0.013**					-0.023	-0.025					-0.018***	-0.014**
					(0.005)	(0.004)					(0.021)	(0.020)					(0.005)	(0.005)
engXteamXRnDlab					0.022						-0.028						0.028*	
					(0.013)						(0.071)						(0.013)	
Constant	0.527***	0.527***	0.525***	0.515***	0.514***	0.514***	0.476***	0.476***	0.475***	0.472***	0.472***	0.472***	0.583***	0.582***	0.579***	0.559***	0.558***	0.558***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 5: Novelty regression - Model C (year×technology- sector fixed effects)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.006***		0.006***	0.015***	0.015***	0.015***	0.012***		0.014***	0.020***	0.020***	0.016***	0.005***		0.004**	0.012***	0.013***	0.014***
	(0.001)		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.003)	(0.003)	(0.001)		(0.001)	(0.003)	(0.003)	(0.002)
team		0.003***	0.003***	0.001	0.001	0.001		0.001	0.001	0.001	0.001	0.000		0.006***	0.005***	0.002	0.002	0.003*
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
engXteam			0.003	-0.001	-0.001				-0.013*	-0.024**	-0.024**				0.005	0.011	0.011	
			(0.003)	(0.005)	(0.005)				(0.006)	(0.007)	(0.007)				(0.003)	(0.007)	(0.007)	
firm				0.006***	0.003***	0.003***				0.002**	0.001	0.001				0.008***	0.005***	0.005***
				(0.000)	(0.000)	(0.000)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
engXfirm				-0.015***	-0.014***	-0.013***				-0.016***	-0.015**	-0.011*				-0.014***	-0.012***	-0.014***
				(0.002)	(0.003)	(0.002)				(0.004)	(0.005)	(0.004)				(0.003)	(0.003)	(0.003)
teamXfirm				0.005***	0.005**	0.005***				0.003	0.003	0.003				0.004*	0.004*	0.004
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)
engXteamXfirm				0.004	0.002					0.027*	0.029*					-0.008	-0.011	
				(0.006)	(0.006)					(0.012)	(0.012)					(0.007)	(0.008)	
RnDlab					0.013***	0.013***					0.009***	0.009***					0.013***	0.012***
					(0.001)	(0.001)					(0.002)	(0.002)					(0.001)	(0.001)
engXRnDlab					-0.009**	-0.007**					-0.018	-0.020					-0.007*	-0.006*
					(0.003)	(0.002)					(0.011)	(0.010)					(0.003)	(0.003)
teamXRnDlab					-0.001	0.000					0.006	0.005					-0.003	-0.001
					(0.002)	(0.002)					(0.009)	(0.008)					(0.003)	(0.002)
engXteamXRnDlab					0.008						-0.016						0.010	
					(0.006)						(0.029)						(0.007)	
Constant	0.054***	0.054***	0.054***	0.052***	0.052***	0.052***	0.044***	0.045***	0.044***	0.044***	0.044***	0.044***	0.065***	0.064***	0.064***	0.060***	0.060***	0.060***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.033***		0.032***	0.059***	0.061***	0.059***	0.050***		0.051***	0.075***	0.075***	0.071***	0.029***		0.028***	0.052***	0.054***	0.052***
	(0.002)		(0.002)	(0.004)	(0.004)	(0.004)	(0.005)		(0.005)	(0.007)	(0.007)	(0.006)	(0.002)		(0.003)	(0.005)	(0.005)	(0.005)
team		0.012***	0.011***	0.006***	0.006***	0.006***		0.004	0.003	0.001	0.001	0.001		0.020***	0.019***	0.016***	0.016***	0.015***
		(0.001)	(0.001)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
engXteam			0.001	-0.013	-0.013				-0.007	-0.031	-0.031				-0.003	-0.008	-0.008	
			(0.005)	(0.011)	(0.011)				(0.014)	(0.018)	(0.018)				(0.006)	(0.013)	(0.013)	
firm				0.020***	0.011***	0.011***				0.005**	0.002	0.002				0.030***	0.018***	0.018***
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.001)	(0.001)	(0.001)
engXfirm				-0.046***	-0.046***	-0.045***				-0.064***	-0.068***	-0.059***				-0.042***	-0.040***	-0.040***
				(0.005)	(0.005)	(0.005)				(0.011)	(0.011)	(0.010)				(0.006)	(0.007)	(0.006)
teamXfirm				0.011***	0.015***	0.015***				0.013*	0.015**	0.017**				0.001	0.006	0.006
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.004)	(0.004)	(0.004)
engXteamXfirm				0.018	0.012					0.055	0.060*					0.012	0.002	
				(0.012)	(0.013)					(0.029)	(0.030)					(0.015)	(0.016)	
RnDlab					0.049***	0.049***					0.044***	0.044***					0.047***	0.046***
					(0.002)	(0.002)					(0.006)	(0.006)					(0.002)	(0.002)
engXRnDlab					-0.019**	-0.014**					0.018	0.012					-0.020**	-0.013*
					(0.006)	(0.005)					(0.027)	(0.025)					(0.006)	(0.006)
teamXRnDlab					-0.019***	-0.015**					-0.021	-0.024					-0.020***	-0.015**
					(0.005)	(0.005)					(0.021)	(0.020)					(0.005)	(0.005)
engXteamXRnDlab					0.027*						-0.043						0.033*	
					(0.014)						(0.071)						(0.014)	
Constant	0.528***	0.528***	0.527***	0.521***	0.520***	0.520***	0.476***	0.476***	0.475***	0.475***	0.475***	0.475***	0.583***	0.583***	0.581***	0.568***	0.567***	0.567***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 6: Novelty regression - Model A* (baseline)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.005***		0.009***	0.001		0.001	0.013***		0.006***	0.015***		0.020***	-0.001**		0.004***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.014***	0.016***		0.002***	0.002**		0.020***	0.019***		0.020***	0.020***		-0.008***	-0.007***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			-0.011***			0.003			0.008***			-0.010***			-0.004***
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.063***	0.057***	0.056***	0.045***	0.044***	0.044***	0.063***	0.055***	0.055***	0.091***	0.081***	0.079***	0.061***	0.066***	0.066***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.114***		0.056***	0.005.		0.002	0.048***		0.027***	0.044***		0.039***	0.019***		0.019***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.152***	0.131***		0.021***	0.020***		0.074***	0.070***		0.082***	0.080***		0.030***	0.028***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			0.017***			0.016**			0.018***			-0.014***			-0.008***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.630***	0.577***	0.570***	0.476***	0.473***	0.473***	0.580***	0.550***	0.548***	0.726***	0.683***	0.678***	0.769***	0.756***	0.752***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 7: Novelty regression - Model B* (year fixed effects)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.007***		0.008***	0.001		0.000	0.008***		0.004**	0.016***		0.020***	0.004***		0.008***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.007***	0.007***		0.005***	0.005***		0.013***	0.012***		0.021***	0.021***		-0.006***	-0.008***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			-0.004***			0.003			0.005*			-0.010***			-0.003*
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.063***	0.061***	0.060***	0.045***	0.044***	0.044***	0.064***	0.059***	0.058***	0.091***	0.080***	0.078***	0.059***	0.066***	0.064***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.025***		0.019***	0.006**		0.003	0.031***		0.021***	0.040***		0.037***	0.023***		0.022***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.044***	0.042***		0.018***	0.016***		0.048***	0.046***		0.078***	0.076***		0.031***	0.028***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			-0.002			0.015**			0.007*			-0.014***			-0.007***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.653***	0.635***	0.632***	0.476***	0.473***	0.473***	0.582***	0.563***	0.561***	0.727***	0.686***	0.681***	0.768***	0.755***	0.750***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 8: Novelty regression - Model C* (year×technology- sector fixed effects)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	-0.000		0.007***	0.001		0.000	0.010***		0.004**	0.010***		0.015***	-0.001**		0.005***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.007***	0.010***		-0.001*	-0.002*		0.014***	0.013***		0.009***	0.010***		-0.010***	-0.009***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			-0.013***			0.003			0.006***			-0.009***			-0.005***
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.065***	0.061***	0.060***	0.045***	0.045***	0.045***	0.064***	0.058***	0.058***	0.092***	0.088***	0.086***	0.061***	0.068***	0.067***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.082***		0.048***	0.003		0.000	0.033***		0.022***	0.029***		0.028***	0.017**		0.019***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.114***	0.101***		0.008***	0.007***		0.053***	0.051***		0.054***	0.053***		0.020***	0.018***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			0.009***			0.016**			0.009*			-0.010**			-0.007***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.638***	0.597***	0.591***	0.476***	0.475***	0.475***	0.581***	0.560***	0.558***	0.730***	0.701***	0.697***	0.770***	0.763***	0.759***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table 9: Model D* (baseline, without non-firm organizations)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.005***		0.009***	0.001		0.001	0.013***		0.006***	0.015***		0.020***	-0.001***		0.004***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.014***	0.016***		0.002***	0.002**		0.020***	0.019***		0.020***	0.020***		-0.008***	-0.007***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			-0.012***			0.003			0.008***			-0.010***			-0.005***
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.063***	0.057***	0.056***	0.045***	0.044***	0.044***	0.063***	0.055***	0.055***	0.091***	0.081***	0.079***	0.061***	0.066***	0.066***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3368109	3368109	3368109	779970	779970	779970	728534	728534	728534	671112	671112	671112	1188493	1188493	1188493

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1975			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.114***		0.056***	0.005.		0.002	0.048***		0.027***	0.044***		0.039***	0.019***		0.019***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.151***	0.130***		0.021***	0.020***		0.074***	0.070***		0.082***	0.080***		0.030***	0.028***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
teamXfirm			0.017***			0.016**			0.018***			-0.014***			-0.009***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.629***	0.577***	0.570***	0.476***	0.473***	0.473***	0.580***	0.550***	0.548***	0.726***	0.683***	0.678***	0.769***	0.756***	0.752***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3368109	3368109	3368109	779970	779970	779970	728534	728534	728534	671112	671112	671112	1188493	1188493	1188493

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

  - The finding holds for novelty defined at the 3-digit level as well as at the 6-digit level
  - Engineers do not generate more novelty inside than outside firms or labs. However, it should be noted that relatively few engineers file patents in a stand-alone capacity (see Fig. 10): in the period 1856-1945, firms and labs account for 72% of engineers and 76% of teams that include engineers. Both shares rise over time, reaching over 80% by the 1920s.

- Teams
  - Teams tend to generate more novel combinations than solo inventors
  - This finding only emerges in the later (1920-1945) period. Before the 1920s, teams don't create more novelty than solo inventors.
  - At the 6-digit level, there is weak evidence that teams create more novelty than solo inventors.
  - The engineering and team effects seem to operate independently, with little evidence of interactions between them.
  - The enhanced novelty of team output seems to be enhanced in teams that patent for firms (with or without industrial research labs)

- Fixed effects
  - Adding year fixed effects has relatively little consequences. This suggests that observed associations between inputs and coordination mechanisms on the one hand and novelty generation on the other are not due to spuriously correlated time-trends.
  - Adding year $\times$ technology fixed effects does not noticeably change results. This suggests that the novelty creating effects of system 2 operate also when comparing patents within technological domains.

Our main focus has been on engineers and teams as labor inputs into the invention process and how these labor inputs are enhanced by new coordination mechanisms. The interaction effects complicate drawing conclusions from the regression tables.[11] To facilitate this, we summarize the main parameters of interest in Figs. 11-16. These figures show how much greater the likelihood of a new combination is on patents by engineers or teams, depending on whether these engineers or teams work for firms or research labs. The omitted category against which these effects are compared consists of solo inventors, who are not engineers and who patent outside firms and labs.

[11] For instance, to calculate the novelty effect of a lab-based team of non-engineers in model 6 of Table 3, we would need to add up the 2nd, 4th, 6th and 8th coefficients and calculate the standard error of this sum.

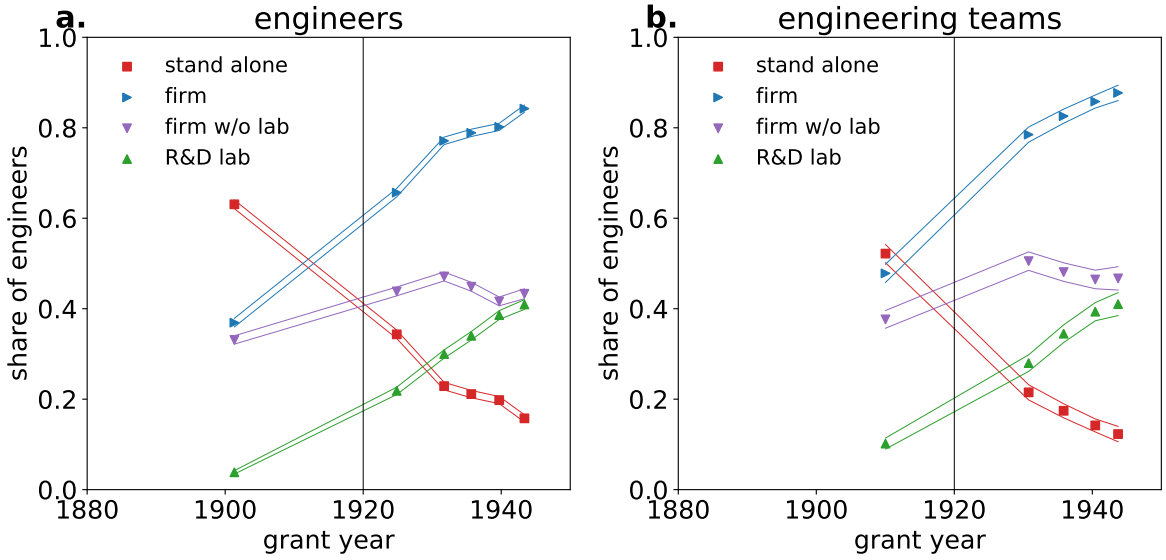


Figure 10: Workplaces of engineers and engineering teams. **a**: Share of inventors with engineering occupations who work as stand-alone inventors (red), for firms (blue), firms without research labs (purple), or firms with research labs (green). **b** Share of team patents that include at least one engineer among the inventors that are filed as stand-alone patents (red), for firms (blue), firms without research labs (purple), or firms with research labs (green).

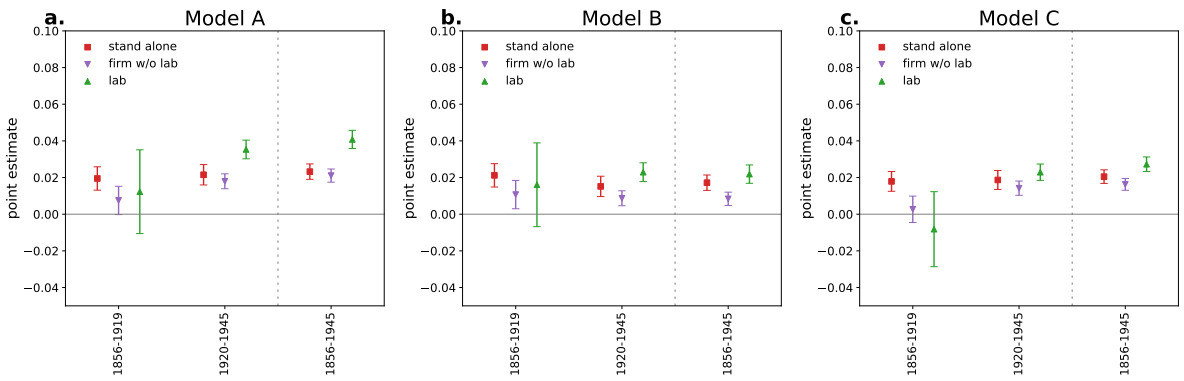


Figure 11: Summary of regression analyses (3-digit novelty): *Engineer effects.* Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. Vertical spikes display 95% confidence intervals.

These figures clearly show that the novelty generation by teams is elevated when these teams patent for firms, and even more so when they do so for industrial research labs (Figs. 13 and 14). Moreover, Figs. 15 and 16 show that the reversal of this effect starts in the 1945-1969 period and fully materializes in the 1975-2000 period. Finally, the greater novelty generation by engineers is only elevated in industrial research labs, and here, this effect is to some extent due to the fact that these labs tend to operate more in technology fields that display a greater propensity to exhibit novel technological combinations (Figs. 11 and 12). These general results are robust across model specifications. The fact that even when controlling for year effects that are allowed to differ by technological domain we see similar effects of engineers, teams, firms and labs on novelty shows that the novelty generating aspects of system 2 are also present when comparing patents within such domains.

The figure highlights four general findings. First, both inputs of system 2, engineers and teams, are associated with elevated technological novelty of their inventions. It is important to note that this holds even before the take-off of system 2. In other words, it is not the case that teams and engineers start becoming more creative in the 1920s, but rather that in the 1920s, they quickly become more dominant. Second, although engineers create more novel patents than non-engineers, this is not enhanced by the stable work environments created by firms or industrial research labs. Third, and in contrast to the latter finding, teams – which create more novel inventions than solo inventors – see their novelty-creating capacity enhanced if they work for firms, and especially for firms with industrial research labs. Fourth, the increased novelty associated with team patents fades over time, peaking in the period immediately after World War II, but dropping precipitously in the last quarter of the 20th century. What is more, in this final period,

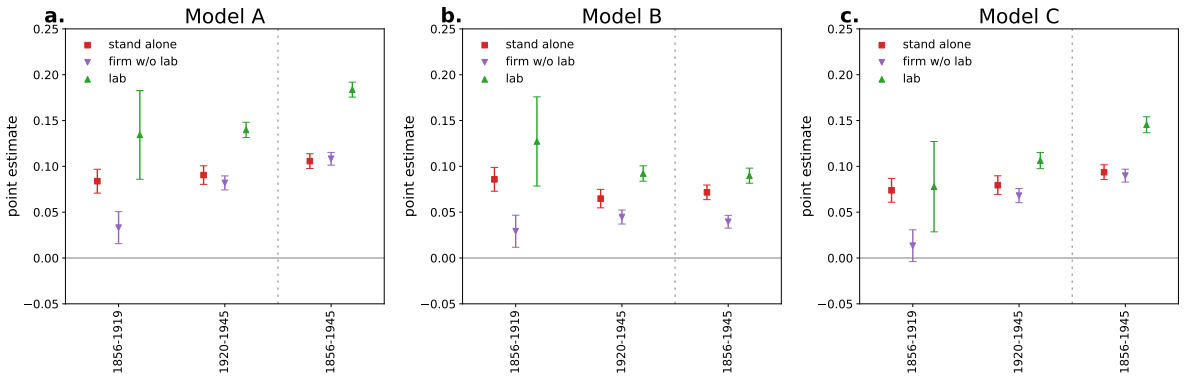


Figure 12: Summary of regression analyses (6-digit novelty). ***Engineer effects*** Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. Vertical spikes display 95% confidence intervals.

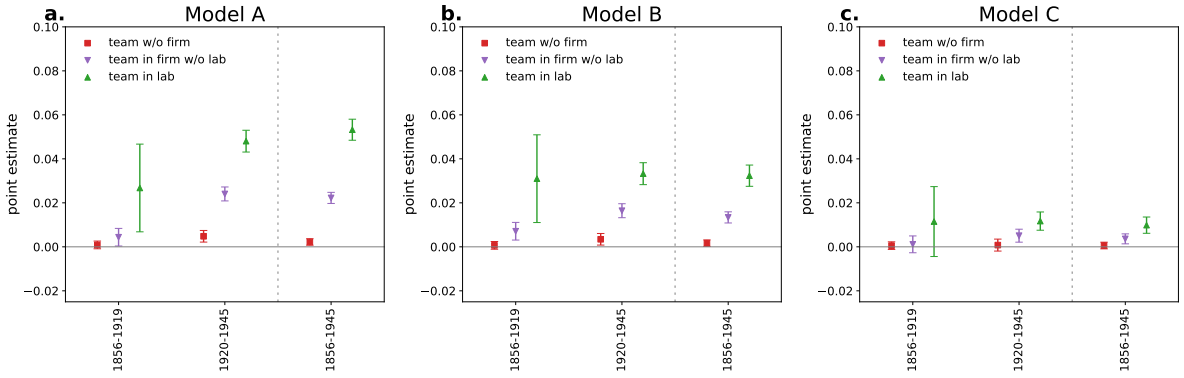


Figure 13: Summary of regression analyses (3-digit novelty): ***Team effects.*** Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. Vertical spikes display 95% confidence intervals.

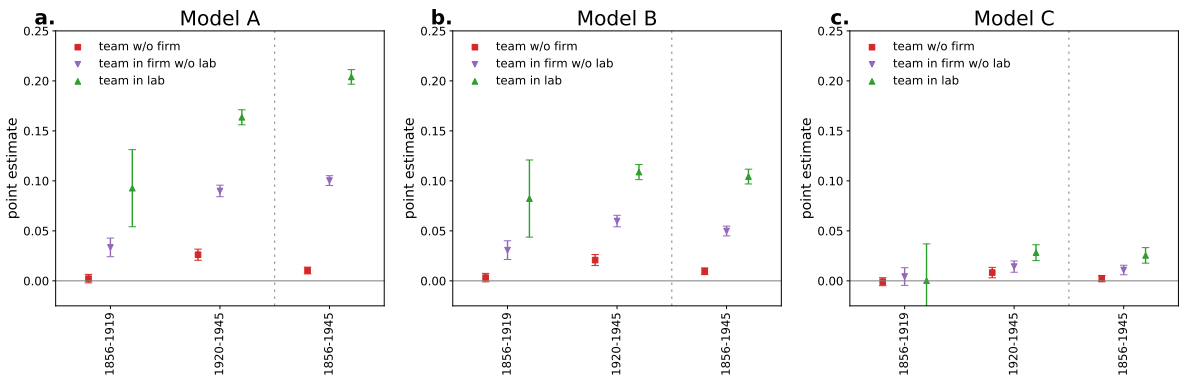


Figure 14: Summary of regression analyses (6-digit novelty). ***Team effects*** Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. Vertical spikes display 95% confidence intervals.

firms stop enhancing the novelty-creating capacity of teams. On the contrary, in this period, firm-based teams produce less radical novelty than stand-alone teams. In fact, the same holds for solo inventors: in the period 1975-2000, solo inventors outside firms produce novel combinations at a higher rate than do solo inventors inside firms. However, these latter findings only hold for radically new combinations, not for the incremental novelty of 6-digit technology class combinations. Here, firm-based patents are associated with slightly more novel combinations than stand-alone patents, regardless whether we look at solo inventors or teams. However, compared to earlier periods, the positive effect of firms on novelty-creation weakens considerably.

A.4.2 Novelty by technology class

Fig. 17 repeats the analysis of Fig. 1a separately for each of six broad technology classes in the NBER classification. Although the share of radically and incrementally novel patents differs, temporal patterns are remarkably similar across these classes.

A.4.3 Spatial distribution of system 1 and 2

Figure 18 shows that system 1 is concentrated in the east of the US over time, and highlights the emergence between 1920 and 1945 of Silicon Valley (San Francisco and San Jose / Santa Clara). System 1 is mainly concentrated in the western part of the US, particularly so in the cities of the Rust Belt region (Chicago, Pittsburgh, Detroit, Cleveland, Buffalo, Milwaukee, Akron, St. Louis, Youngstown) as well as those on the east coast (Boston, New York).

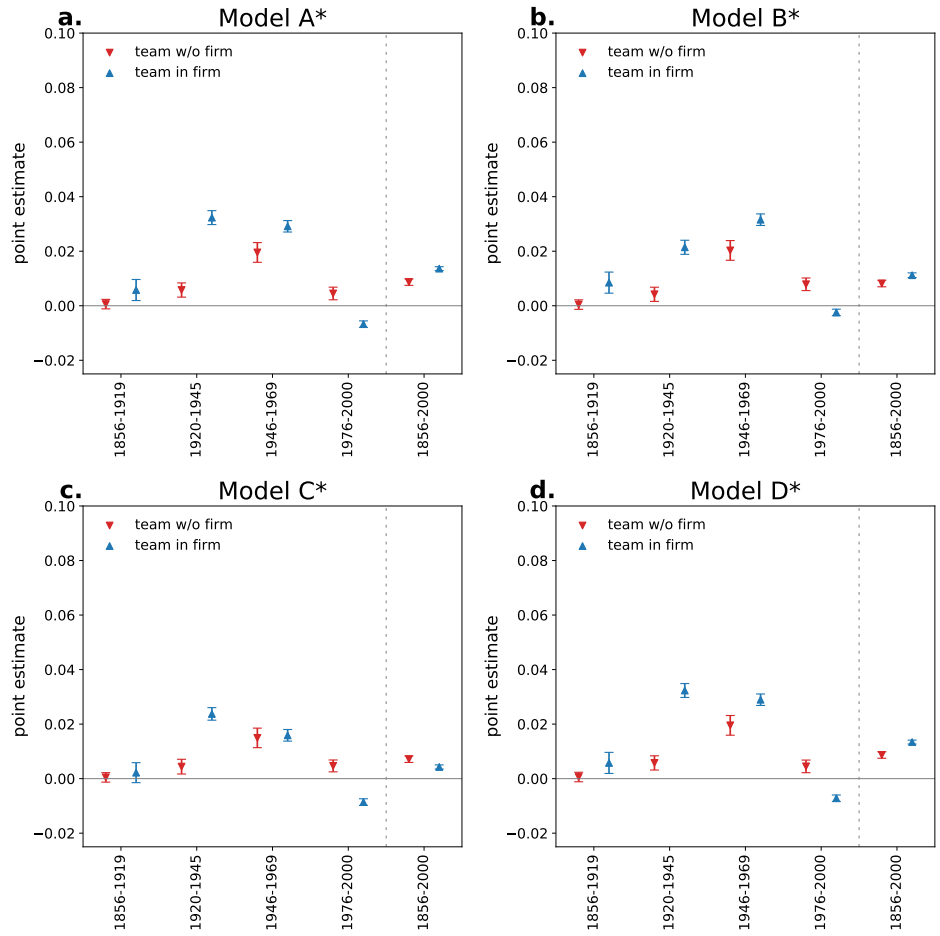


Figure 15: Summary of regression analyses (3-digit novelty). Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. **c**: Model D*: baseline, patents assigned to non-firm organization dropped. Vertical spikes display 95% confidence intervals.

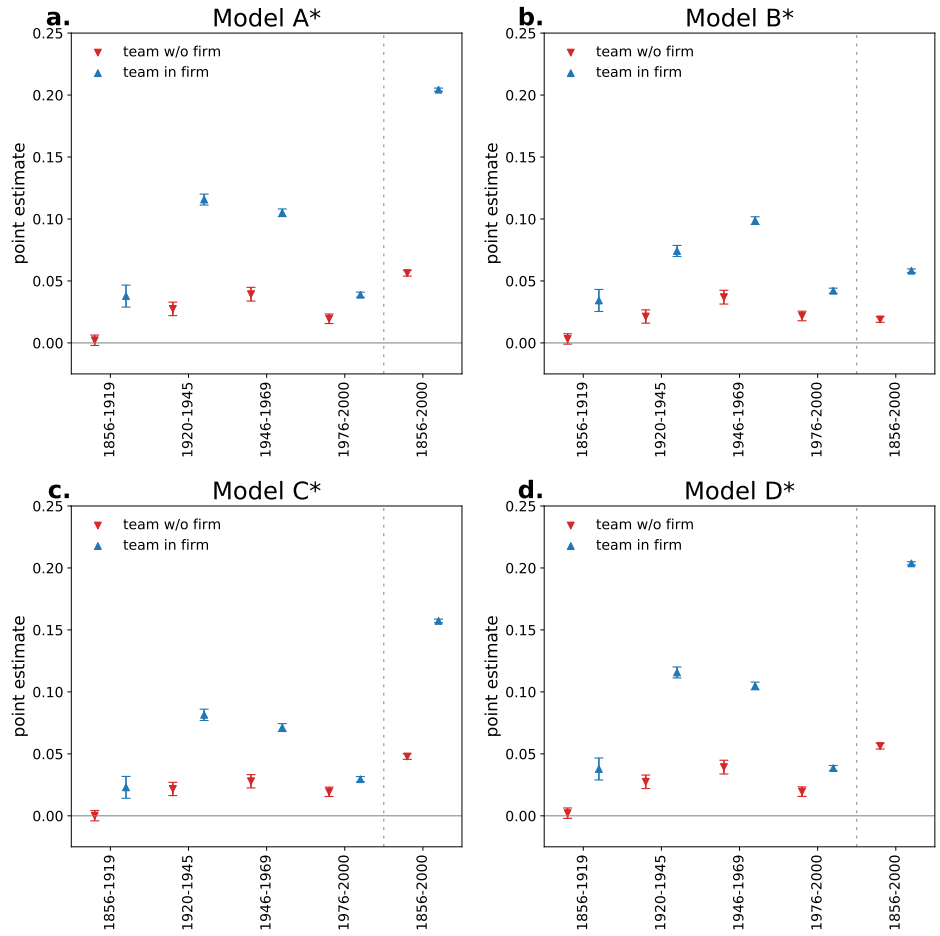


Figure 16: Summary of regression analyses (6-digit novelty). Omitted category is patents by solo inventors outside firms. **a**: Model A*: baseline. **b**: Model B*: baseline + year fixed effects. **c**: Model C*: baseline + sector × year fixed effects. **c**: Model D*: baseline, patents assigned to non-firm organization dropped. Vertical spikes display 95% confidence intervals.

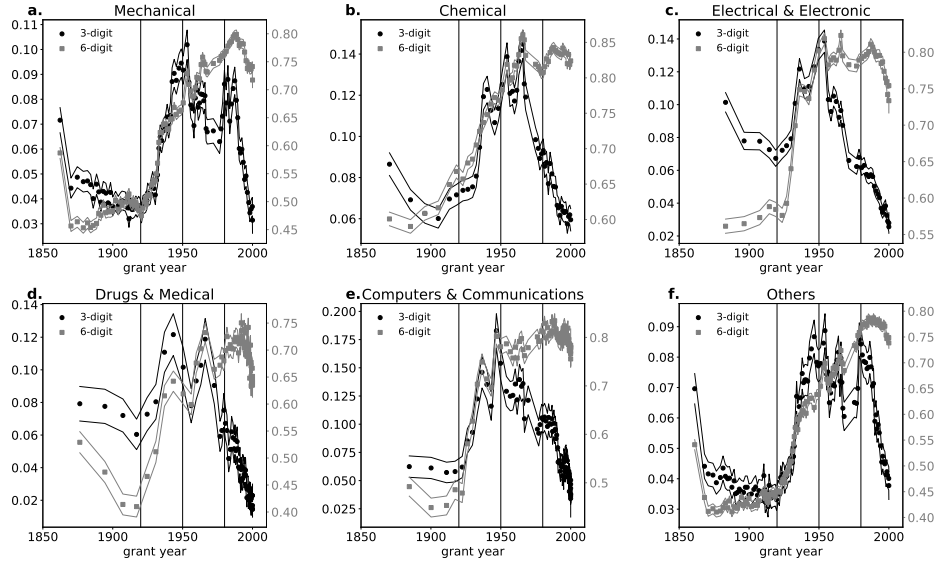


Figure 17: New combinations by broad technology class. Each graph shows the share of new 3-digit (black) and 6-digit (gray) technologies by the six broad primary classes defined in [32].

A.4.4 Team composition

Over time, teams become more homogeneous in terms of the occupational backgrounds of team members (Fig. 19a and b), as well as their age (Fig. 19c), although the age difference starts rising rapidly again in the 1930s, a phenomenon driven by firm-patents. Closer inspection of the occupations that are combined in teams seem to suggest that these patterns reflect a move from master-apprentice teams in the 19th century to more egalitarian teams in the 20th century.

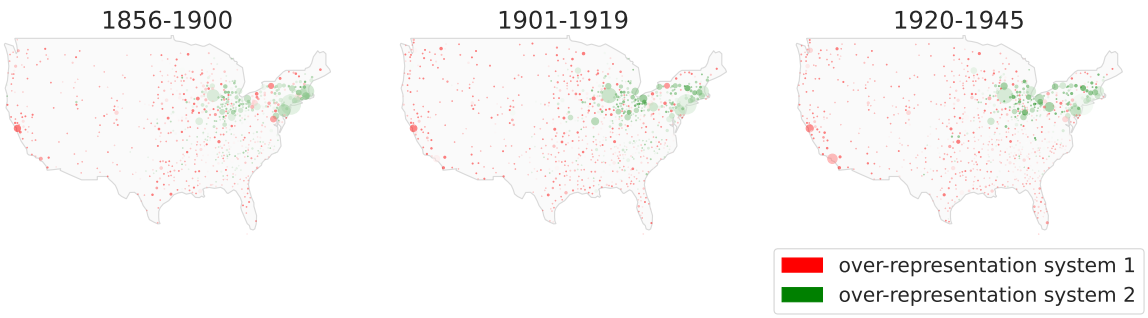
Differences in occupational and age homogeneity between firm- and lab-based teams and stand-alone teams are somewhat erratic. However, there are pronounced differences between these types of teams when it comes to the distance over which inventors collaborate. This is shown in Fig. 19d.

Fig. 19d was created by estimating gravity models - following [41] - that try to predict the size of collaboration flows X between all pairs of cities i and j until 1950:

$$X_{ij} = \exp\left[\gamma_i + \eta_j + \beta Z_{ij}\right] + \varepsilon_{ij} \tag{4}$$

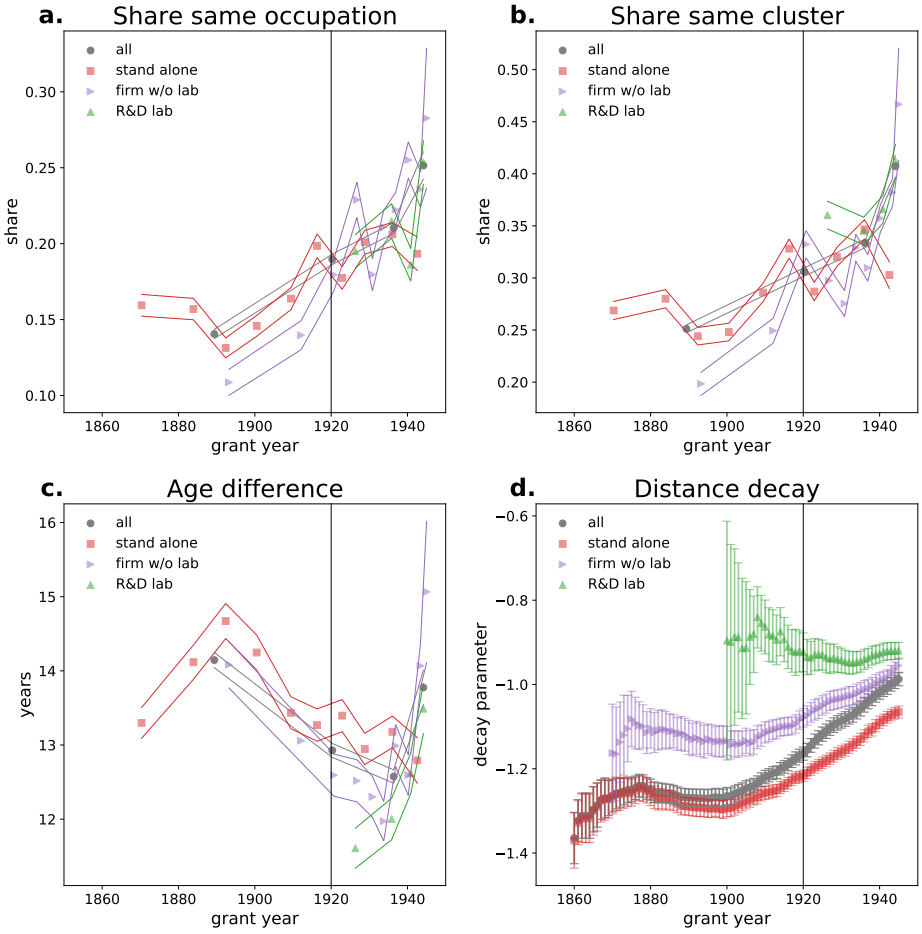
where γ_i and η_j are city fixed effects and Z_{ij} represents the bilateral determinants of collaboration, in our case the logarithm of the Haversine ('as the crow flies') distance. To estimate the parameters we rely on the Pseudo-Poisson Maximum Likelihood (PPML) estimator by [42] which yields robust estimates of the parameters in gravity models.

Figure 18: Over-representation of system 1 and 2 across US cities



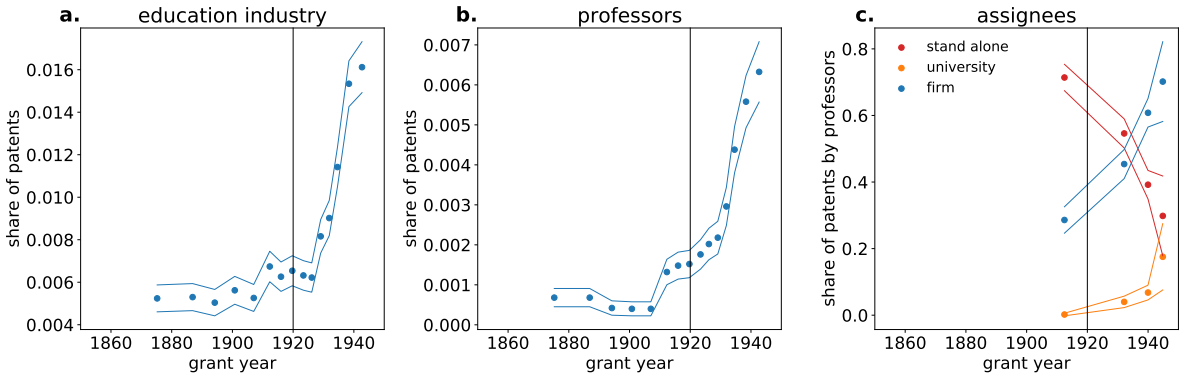
Graphs show which system is over-represented in each of 933 US cities, where over-representation is defined as the ratio of the city's share of patents in one system over the city's patents in the other system. Green tones mean that system 2 is over-represented in the city, red system 1. The darker the tone, the greater the over-representation. Marker sizes depict the share of patents in the city of the total count in the US.

Figure 19: Distances between inventors on team patents



a: share of inventor pairs with the same detailed occupation (256 number of classes) on a team patent and **b** broad flow-based clusters (14 classes, green) on a team patent. **c**: average age difference between inventors on a team patent. **d**: distance decay parameter for collaboration flows between US cities.

Figure 20: Academic patents



a: Share of inventors that list the education industry as their industry; **b**: share of inventors that list professor as their occupation; **c**: share of patents with at least one inventor that list professor as their occupation who assign their patent to a firm (blue), a university (red) or who don't assign their patent to any organization (red).

Next, we plot how the estimated distance decay parameter, $\hat{\delta}_\theta$ for collaborations in team type θ change over time. The less negative this parameter is, the less distance constrains collaboration. Panel d of Fig. 19 shows that long-distance collaborations were much more constrained in the 19th than in the 20th century. From the year 1900, collaborations seem to overcome longer distances more easily. This shift is driven by two forces. First, stand-alone teams and teams in firms without labs are decreasingly constrained by distance. Second, more and more patents are filed by lab-based teams, which are much less constrained by distance throughout their existence. This suggests that labs were able to help teams overcome frictions in long-distance collaborations.

A.4.5 University patents

As invention becomes progressively the domain of engineers, do we also see a greater involvement of the universities? To analyze this, in Fig. 20 we plot the share of inventors who list "education industry" as their industry or "professor" as their occupation. Academic patents are very rare until the 1910s and then see a rapid rise in the 1920s. However, only in the 1940s do universities actively start patenting themselves and showing up as assignees in the patent records.