What Happens When Inventors Collaborate? Evidence from 19th-Century French Patents

Latest draft

Youssouf Merouani - Lund University

youssouf.merouani@ekh.lu.se

Preliminary Draft

2026-01-01

Abstract

Collaboration was rare in nineteenth-century France, yet when inventors moved from solo work to co-patenting it reshaped their inventive activity in ways distinct from modern teamwork. Using the full universe of French patents from 1791 to 1900, text-based measures of novelty and influence derived from patent titles, and within-inventor event studies around each inventor's first collaboration, I show that collaboration did not increase productivity or originality and produced only a short-lived rise in influence. Its durable effect was a large and persistent expansion in technological scope: first collaboration shifted roughly eleven percentage points of output outside an inventor's historical home class. Network evidence indicates that this diversification was driven by partners from different technology classes and by repeated ties that had already crossed a technological boundary. Women, who were scarce and often entered invention through family ties, benefited disproportionately from these cross-class and repeated outside class collaborations, achieving substantially larger shifts in scope than men. The results suggest that in this historical setting collaboration functioned primarily as a channel for knowledge access and boundary-crossing rather than as a generator of more novel or more productive ideas.

1 Introduction

Modern innovation is overwhelmingly collaborative. In science and patenting, teams have come to dominate the upper tail of impact and are often credited with producing the most influential and technically complex work (Jones, 2009; Wuchty et al., 2007). Teams mix diverse knowledge (Uzzi et al., 2013), generate fewer but more refined recombinations, and, depending on size, either disrupt or deepen existing lines of work (Wu et al., 2019). Yet, we know far less about how collaboration operated in earlier technological regimes, where knowledge was thinner, much technical know-how remained tacit rather than written down, and institutional constraints shaped who could work with whom.

Nineteenth-century France provides a contrast to the modern settings. The French patent system, grounded in natural-rights ideology, has been described both as open and accessible (Galvez-Behar, 2019; Merouani and Perrin, 2026) and as oligarchic, with high fees and secrecy that favored elites (Khan, 2020). Collaboration was rare: most inventors worked alone, and the co-inventor network consisted largely of isolates and dyads. If collaboration is as central to innovation as modern evidence suggests, why was it so limited? And when it did occur, what did it change?

This paper uses the full universe of French patents from 1791 to 1900 to examine how moving from solo work to collaboration altered inventive behavior in this earlier, sparsely networked system. Because historical patents lack citations, I develop text-based measures of novelty and influence from patent titles and pair them with within-inventor event studies around each inventor's first collaboration and a reconstruction of the complete collaboration network. This allows me to study both the immediate effects of collaboration and the network mechanisms—structural holes, cohesion,

and boundary-spanning—that the literature associates with access to non-redundant knowledge (Burt, 2004; Fleming et al., 2007b; Reagans and McEvily, 2003).

Three findings emerge. First, collaboration did not raise productivity or novelty. Output spikes only mechanically in the year of the joint patent, novelty remains flat, and influence rises modestly before fading—patterns far removed from the large citation and novelty gains documented in modern teams. Second, collaboration produced a substantial, persistent shift in technological scope: inventors redirected roughly eleven percentage points of their output outside their historical home class at first collaboration. Third, this diversification was not uniform. It was strongest when inventors worked with collaborators from different technology classes and when they repeated ties with partners who had already helped them cross a technological boundary. Peer exposure mattered little. Women—scarce in this system and often reliant on family ties for access—benefited disproportionately from such cross-class and repeated outside ties, consistent with research showing that women face higher thresholds of trust and legitimacy in networks rich in structural holes (Ling et al., 2025; Meng, 2016).

These results suggest that collaboration in nineteenth-century France functioned primarily as a mechanism for knowledge access and dissemination across fields, not as a means of producing more original ideas or becoming more productive. The dispersed network structure, with few brokers, no dominant hubs, and many small components: matches this interpretation and aligns more closely with accounts of the French patent system as broadly open than with narratives of elite control. The paper contributes methodologically by introducing text-based measures suitable for historical innovation contexts and substantively by revealing how collaborative ties shifted the technological direction, rather than the volume or originality, of inventive work in a formative period of industrialization.

2 Literature framework

2.1 Patenting and France's economic context in the nineteenth-century

For decades, historians saw French industrialization as a failure. Clapham (1921) and Landes (1949; 1969) blamed conservative social structures and stubborn attachment to small-scale craft production. Roehl (1976) and later O'Brien and Keyder (1978) flipped the narrative: France did not lag; it specialized. Where Britain mass-produced, France made quality goods, especially in textiles and luxury items (Walton, 1992). The numbers support specialization. Textiles employed 40 percent of France's industrial workers in 1860. Textiles and food together generated 62 percent of industrial output (Merouani and Perrin, 2026). This was an economy built on craft skill and niche production, not mass manufacturing. Collaboration networks, if they existed, had to span fragmented workshops and dispersed knowledge.

France made patents a natural right on January 7, 1791 (Marchal, 2009). Enlightenment principles drove the choice: inventors owned their ideas, and the state could not censor them by judging novelty in advance (Galvez-Behar, 2019). Five-year patents cost 300 livres plus a 50-livre fee; ten years cost 800; fifteen cost 1,500. Those sums excluded most artisans and favored inventors with money or backers (Khan, 2020). The law added another burden: use your invention within two years or lose your claim (Galvez-Behar 2011). That clause raised the stakes beyond filing cost.

No formal examination did not mean the absence of scrutiny. The Comité consultatif des Arts et des Manufactures reviewed applications informally, offering opinions that could shape outcomes and adding a degree of discretion to what was intended as an open system (Baudry, 2015, 2019; Hirsch and Minard, 1998). Patent details stayed locked in manuscript at local offices, rarely printed, which choked off knowledge spillovers (Galvez-Behar, 2019; Khan and Sokoloff, 2008). Worse, the law granted "import patents" to anyone bringing foreign inventions to France first, rewarding middlemen over creators (Emptoz and Marchal, 2002; Galvez-Behar, 2019). Import patents moved technology across borders (Nuvolari et al., 2023) but corrupted the link between invention and reward.

July 5, 1844, brought change. The new law clarified property rights, patents now explicitly gave inventors exclusive commercial control (Article 1), and killed import patents to favor original work over borrowed ideas (Galvez-Behar, 2019). Foreigners could file on equal terms, and annual

installment payments of 100 livre cut upfront costs, mimicking the accessible American model (Khan, 2005). The overall fees also became relatively cheaper for the longest duration. Courts, not bureaucrats, judged validity. Patents still carried no government guarantee (Article 33), which preserved ideological purity but left inventors vulnerable to litigation (Galvez-Behar, 2019). The reforms coincide with increase in levels. Annual patents jumped from 1,200 in the early 1840s to 5,800 by the late 1850s (Merouani and Perrin, 2026).

Khan (Khan, 2005, 2020) calls the French system oligarchic: high fees, secrecy, and rent-seeking locked out common inventors and protected elites. Galvez-Behar (Galvez-Behar, 2019) counters that natural-rights ideology made French patents more democratic than Khan admits. Merouani and Perrin (2026) agree with the latter view. They find that among women patentees with known occupations, 31 percent held medium-skill jobs and 25 percent came from the bottom: seamstresses, tailors, unskilled laborers. France issued more patents per capita than the United States during 1840–1860 and matched American rates of one-time patenting, a key accessibility marker (Nuvolari et al., 2023). High fees mattered, but they did not close the door; especially so after the 1844 reform.

2.2 Why collaborate? Complexity in knowledge production and collaboration

As knowledge accumulates and individual scope narrows, collaboration rises to cover missing expertise. Jones (2009), using U.S. patents (1975–1999) and scientific publications (1955–2000), finds that age at first invention rises by about 0.6 years per decade, team size grows roughly 17 percent per decade, and specialization deepens, measured by a ~6 percent per decade decline in switching across technological classes between consecutive patents. Across science, social science, arts and humanities, and patents, teams increasingly occupy the top of the citation distribution by the late 1990s, displacing solos at the upper tail (Wuchty et al., 2007).

Complexity is uneven across fields, and collaboration does not automatically raise individual productivity. Teams in patenting are larger within fields where knowledge is deeper (Jones, 2009). At the macro level, sustaining semiconductor progress requires far more researchers: keeping Moore's Law on pace involves roughly eighteen times the research effort relative to the early 1970s, with research productivity declining about seven percent per year (Bloom et al., 2020). In research output, collaboration correlates with higher total publication counts but not with higher per-author productivity (Lee and Bozeman, 2005), echoing the macro evidence that collective invention expands while returns per inventor fall. At a non-individual actor level, evidence from Sweden (1970–2021) finds that collaboration is generally associated with higher subsequent innovation, though its benefits hinge on partner diversity rather than on sectoral focus (Kreutzer and Taalbi, 2025). These patterns document persistent pressure toward teamwork and specialization.

The nature of recombination also differs between solo and team work. Uzzi et al. (2013) show that the most influential scientific papers mix a small share of atypical, cross-domain combinations into otherwise conventional work, and that teams are markedly more likely than solo authors to achieve this balance. Teams introduce novelty without departing too far from established knowledge frames, whereas solo authors often lack the anchoring conventionality that makes novel work influential. Chan et al. (2021) show that collaboration generally increases the likelihood of breakthrough inventions for functional or technological patents, but this advantage disappears for design patents, whose interdependent elements cannot easily be divided among team members. Wu et al. (2019) differentiate between the sizes of teams across the production of scientific papers, patents, and software and find that smaller teams smaller teams tend to disrupt science and technology with new ideas and opportunities, while larger teams tend to develop existing ones. Patent data also show that collaboration reshapes both tails of the impact distribution: compared with unaffiliated solo inventors, patents produced by teams or within firms are more likely to become highly cited and less likely to be ignored. The combination of teamwork and firm affiliation roughly doubles the chance of producing a breakthrough, with the advantage linked partly to teams' diverse experience and external networks (Singh and Fleming, 2009).

Rising knowledge complexity, specialization, and the division of problem solving go hand in hand. In knowledge hierarchies, routine problems are solved at lower levels and exceptional problems escalate to specialists (Garicano, 2000). In invention, collaborators supply non-overlapping capabilities, such

as conceptual design, methodological execution, implementation, and commercialization, which are costly for one inventor to accumulate as knowledge becomes more complex. Solo breakthroughs persist when problems are simple or when matching frictions are high, so collaboration is conditional on complexity. Entry into collaboration should coincide with exposure to non-redundant knowledge and a larger feasible problem set; whether such gains persist beyond the first collaboration is an open empirical question, distinct from teams' higher citation impact.

2.3 Networks, ideas and knowledge production

2.3.1 Network position, brokerage, and boundary-spanning

Positions in a network shape both where new ideas come from and whether those ideas spread. Brokers who span "structural holes" enjoy a vision advantage: by bridging otherwise disconnected groups, they see alternative frames and can synthesize options others miss (Burt, 2004). Classic weak-tie logic explains part of this advantage: weak ties connect dissimilar circles and supply non-redundant information that is especially useful for exploration, mobility, and diffusion (Granovetter, 1973). Yet novelty creation and knowledge use rely on different structural ingredients. Cohesion (dense, overlapping ties) builds trust and eases transfer of fine-grained, tacit knowledge, while range (ties into diverse pools) equips actors to translate ideas across audiences; both cohesion and range independently facilitate knowledge transfer beyond the effect of dyadic tie strength (Reagans and McEvily, 2003). At the individual level, weaker ties are generally creativity-enhancing, whereas strong ties are neutral on average; moreover, being central helps creativity only when external ties are limited—pairing high centrality with many outside ties is not optimal (Perry-Smith, 2006).

Evidence from patenting sharpens these mechanisms. In career-level data on U.S. collaborative inventors (1975–2002), collaborating through brokered structures is linked to more "generative creativity", the novel combinations of patent subclasses, but brokered ideas are less likely to be used later by others; a diffusion penalty that cohesion can mitigate (Fleming et al., 2007b). At the regional scale, the widely assumed benefits of "small-world" structure do not materialize once measured directly: clustering per se shows no robust association with innovative output, whereas shorter average path lengths and larger connected components are positively associated with subsequent patenting (Fleming et al., 2007a). Taken together, these findings suggest that what matters is not generic clustering but (i) access to non-redundant collaborators that reduce social distance to diverse knowledge, and (ii) sufficient connectedness to move ideas once generated.

2.3.2 Gender and Networks in Knowledge production

Network structure matters for gender gaps knowledge production. Women's underrepresentation in science and patenting persists to this day despite decades and centuries of rising participation (Hanson, 2010; Jung and Ejermo, 2014; Khan, 2024; Merouani and Perrin, 2026; Rossiter, 1995). Structural holes, typically confer advantages by providing access to diverse, non-redundant information (Burt, 1992, 2004). Yet hole-rich networks can disadvantage women, particularly when they lack legitimacy as insiders (Burt, 1998; Ling et al., 2025). The core mechanism is trust: gatekeepers require stronger signals of competence from outsiders before granting access to resources (Lin, 2001; Meng, 2016). Men access commercialization knowledge through weak ties like former students and acquaintances; women access the same knowledge through channels requiring higher trust, particularly those they directly collaborate with (Ding et al., 2006; Meng, 2016) or for instance family members. Once women secure trusted ties, the return deficit disappears, resources leveraged through high-trust collaboration produce equivalent or superior outcomes compared to men (Meng, 2016). The problem from this network perspective is access structure.

Context determines whether network advantages translate into outcomes. Tokenism theory predicts that raising women's proportional representation reduces bias (Kanter, 2010), but intrusiveness theories predict the opposite: as minority representation grows, perceived competition triggers backlash (Blalock, 1967; Yoder, 1991). Empirical evidence confirms that the effects of structural holes are contingent on women's proportional representation in the field. Structural holes predict citations positively for women in low-representation settings but become negative once women's field representation exceeds certain threshold, consistent with backlash rather than inclusion (Ling et al., 2025). The magnitude of boundary-spanning returns varies with the broader institutional environment.

3 Materials and methods

3.1 Data sources and Inventor-identification

I begin with the French Patent Database, 1791–1900, which records roughly 390,000 patents as patent-level entries (Merouani and Perrin, 2026). Each entry lists one or more inventors associated with a single patent ID, but the database offers no way to determine which inventors are unique across the full corpus. To address this, I reshape the data into an inventor–observation table in which each inventor–patent pairing constitutes a distinct row, yielding about 448,000 observations. I then classify these inventor-observations into four types of entities, each requiring its own record-linking strategy: 406,868 male individual inventors, 6,916 women[1] individual inventors, 8,205 family-firm inventors, and 26,156 non-family firm inventors.

The difficulty of record linking varies across these groups. Firms are the most straightforward, because the samples are smaller and names tend to be stable. Women inventors pose greater challenges: the small sample size belies the complexity introduced by naming conventions, including surname changes with marital status. Male inventors are the hardest, largely because of scale. With more than 400,000 observations, even minor ambiguities, such as missing or very common given names, compound into a substantial linking problem.

3.1.1 Firm inventor-identification

I rely on assumptions based on close inspection of the underlying records, including application documents, signatures, dates, and places, to determine when two firm–inventor observations refer to the same entity. These assumptions are chosen to match the qualitative structure of the data for as many entities as possible. A family firm is never linked to a non-family firm, and vice versa. I distinguish between the two by using naming conventions. Non-family firms usually include generic multilingual organizational markers such as Fabrik, Société, or Company. Family firms often signal kinship ties in their names, for example "& Fils," "& Sons," or "Frères." A firm is assumed to keep the same name over time. Even though a firm may persist across generations of directors, I also assume that it will not survive more than fifty years without producing a new invention. Two observations more than fifty years apart are therefore never linked to the same entity unless a third observation falls within fifty years of both and connects them. In practice, almost no observations reach this threshold. It is included mainly to prevent linking firms across implausibly long periods of inactivity.

In addition to these core assumptions, I apply three restrictions separately in order to construct alternative sets of links. In one case, I require inventors to share the same HISCO code, the historical international classification of occupations. In another, I require their inventions to overlap in technical vocabulary, which I capture by identifying at least one common keyword extracted from the patent title [2]. Throughout the paper, "technology class" refers to the twenty broad technological sectors defined by the Institut national de la propriété industrielle (INPI), such as Textile, Machinery, Medicine and Hygiene, and Clothing. I follow Merouani and Perrin (2026) for the construction and assignment of these classes.

I set up additional rules to avoid implausible links, such as linking within the same patent application, and stabilize linking across patent applications within the same patent family [3]. Taken together, these restrictions yield about 201,000 firm comparison pairs and 51,000 family-firm pairs. When clustered, these pairs resolve into $12,315$ distinct firms and $3,242$ distinct family firms.

3.1.2 Individual inventor-identification

For individuals, the first name is the most important complement to the surname in distinguishing inventors across patents. To enrich the French Patent Database, I therefore fill in missing first names by turning directly to the original patent application documents. For female inventors,

[1] I follow Merouani and Perrin (2026) approach to gender identification, which uses gender-specific honorifics for women: 'madame,' 'mademoiselle,' and 'veuve' found in the manual transcription of patent notes. I extend this by detecting the same honorifics directly in the patent scans, which increases the count of women-linked patents by 10. See Appendix B.

[2] See technical keyword section in Appendix A.1.

[3] See all firm blocking rules in Appendix A.2.1.

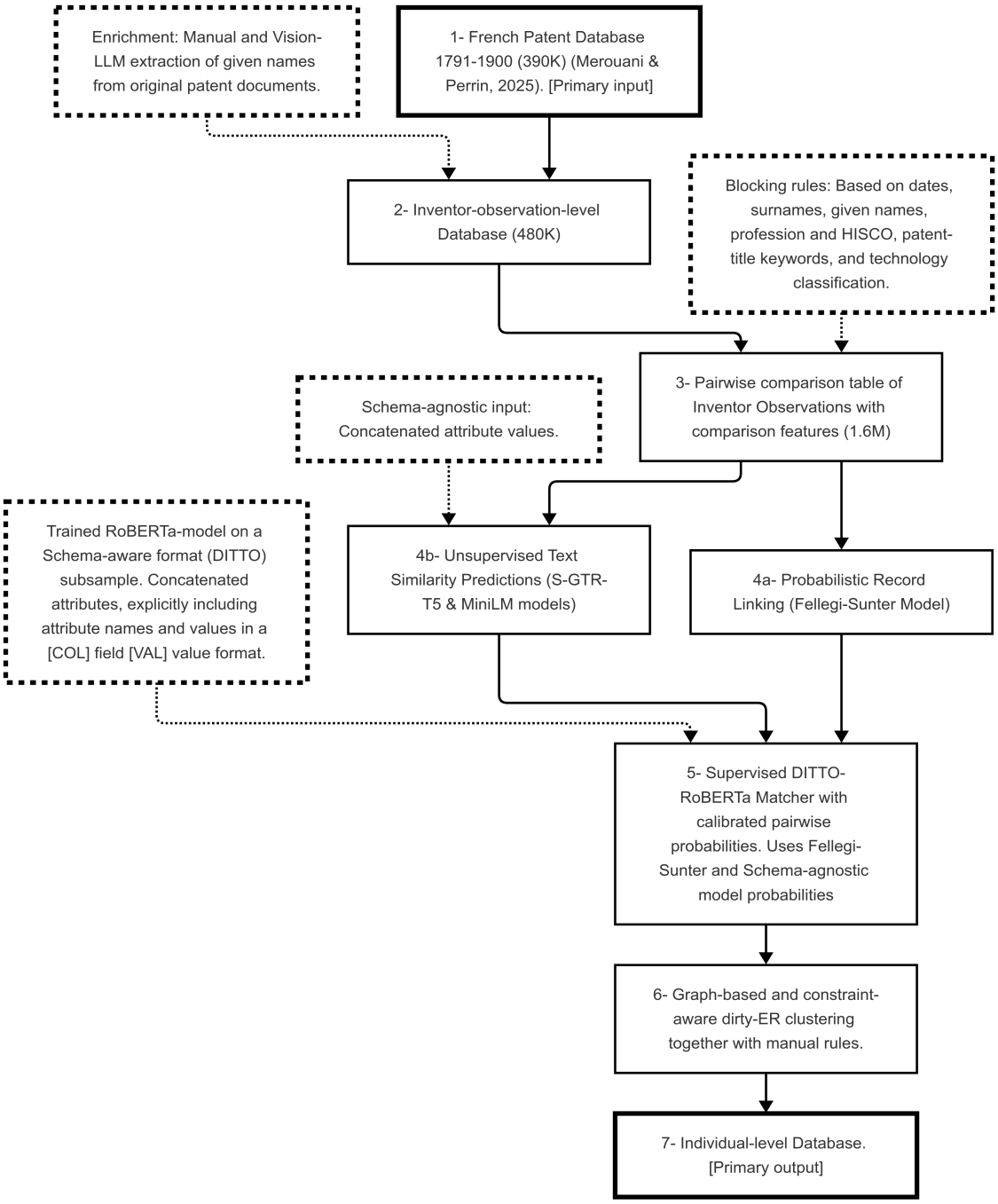


Figure 1: Diagram of the Record Linking process for male individual inventors

I manually transcribe all missing first names. For male inventors, I use an ensemble of three vision-based large language models to extract first names from more than 150,000 handwritten applications that otherwise lacked this information.[4] I then quality-assure and manually clean the transcriptions. When the models do not agree, I resolve the cases myself, focusing especially on the original applications linked to the twenty most frequent surnames. Altogether, I decrease missingness in the first name field from 65 to 33 percent for male inventor-observations.

As with firms, I rely on blocking rules to generate candidate matches: about 13,000 comparison pairs for women inventors and more than 1.5 million for male inventors.[5] The female comparison pairs are then matched using 51 customized rules, each tailored to a specific blocking strategy. These rules, developed through careful manual matching, are designed to be replicable and extendable as the database grows. They draw on features such as surname, first name, marital status, maiden name, occupation, location, year of application, patent identifiers, patent titles, and co-inventors.[6] Once matches are identified, I cluster the pairs into inventor-level entities totalling 4,166 distinct female inventors.

Linking male inventors is the hardest part of the exercise. The pair comparison table holds over 56,000 rows of men with the most frequent surname Martin, and 9,700 of the runner-up "Smith". A single missing first name, or a slight variation like "Jean" versus "Jean-Baptiste," can make one person look like several or several people look like one.

I begin with rules, which resolve about 60 percent of the cases. Some inventors appear on every application within a patent family, leaving little doubt about their identity. Others can be ruled out immediately when their first names or important details conflict. These "must-link" and "cannot-link" rules are cheap and reliable, but they run out of power quickly. The remaining 40 percent of pairs are ambiguous, and here I need more sophisticated tools. I build a four-stage process to handle them, sketched in Figure 1.

The first step is to bring in a probabilistic framework. I use the Fellegi–Sunter model (Enamorado et al., 2019; Winkler, 1988), which assigns a probability that two records belong to the same inventor based on their agreements and disagreements. It accounts for the fact that some features are more informative than others: sharing the surname "Siemens" means more than sharing the surname "Martin." Yet even here the model has limits. It struggles to link the anglicized name of Carl Wilhelm Siemens, Charles William Siemens, without pulling in many false matches. The model is noisy, but it gives a useful baseline.

The second step is to collapse information into text and let modern language models judge how close two records sound to one another. I combine surname, first name, occupation, technology class, co-inventors, year, and keywords from the patent title into a single string for each record. Models like S-GTR-T5 and MiniLM, which have been shown to excel at this kind of similarity task (Zeakis et al., 2025), produce scores that capture subtle connections: "forge master" and "iron works" may not look alike to a rule-based system, but they do to a model trained on language. These schema-agnostic comparisons widen the lens, picking up cases that fixed fields cannot.

The third step is to refine these signals with supervised learning. I draw a training sample of 25,000 pairs, balanced between matches and non-matches, and I deliberately degrade, dropping first names or occupations, so that the training set mirrors the imperfections of the real data. I then fine-tune a RoBERTa-large model using the DITTO framework (Li et al., 2023; Zeakis et al., 2025). Unlike the schema-agnostic approach, DITTO reads the attributes in a structured format, learning which pieces of information matter most. The result is a calibrated match probability that is sharper than either the rules or the unsupervised similarities alone.

The final stage clusters these predictions into inventor identities. The clustering procedure respects the hard rules from the first step: matches that must hold are enforced, and conflicts that cannot hold are excluded. Within these bounds, I build graphs of candidate inventors, add edges when probabilities exceed calibrated thresholds, and merge nodes while respecting realistic limits, such as a maximum career span. Where clusters stretch implausibly far, weak edges are cut to restore plausibility.

[4] See procedure for machine transcription in Appendix C.
[5] See exhaustive list of blocking rules for male inventors in Appendix A.2.3 and female inventors in Appendix A.2.2.
[6] See Appendix A for a list of all features used.

The logic of this sequence is pragmatic. Rules handle the obvious cases. Probabilistic models add a foundation of structure. Language models broaden the net, capturing similarities that rules miss. Supervised learning sharpens the distinctions. Clustering then imposes discipline to transform noisy fragments into coherent inventor entities. Each step has limits, but together they allow me to move from scattered patent filings to a credible reconstruction of 235,988 distinct male inventors from 406,868 observations. A ratio of 1.7 observations per male inventor entity, closely matching the 1.65 observed for the manually linked female inventors.

3.2 Novelty and Influence

Economists have long measured novelty by asking how far a new idea sits from the existing stock of knowledge at the moment it arrives. The classic tools use backward citations and technology classes. Hall, Jaffe, and Trajtenberg's originality index summarizes whether a patent cites prior art concentrated in a narrow set of technology classes or scattered across many (Hall et al., 2001). Jaffe-style proximity treats ideas as points in a high-dimensional space and asks whether a new point lies in a crowded neighborhood or on the fringe (Jaffe, 1986, 1989). The recombinant search tradition, from Fleming through the atypical combinations literature, views novelty as rare pairings of components that are seldom seen together (Fleming, 2001; Uzzi et al., 2013; Weitzman, 1998). The common intuition is simple. An invention is novel when it draws weakly on dense local precedent, especially when it blends ingredients from distant parts of the knowledge map rather than reusing what is already abundant nearby. This framing sits squarely in the knowledge production tradition (Griliches, 1990).

Influence or importance is the forward mirror of that logic. The standard measures follow how often later patents cite a focal patent and how widely those citations spread across technology classes. Forward citations are taken as a noisy but informative count of downstream use, and the generality index asks whether those downstream uses stay inside a technology class or spill broadly (Hall et al., 2001). Work that links citations to economic value shows that influential inventions leave many traces in the subsequent record. Forward citations correlate with firms' market value (Hall et al., 2005), with survey-based and renewal-based value measures (Harhoff et al., 1999; Pakes, 1986; Schankerman and Pakes, 1986), with litigation risk as a proxy for stakes (Lanjouw and Schankerman, 2001), and with the resale and reassignment of patents in secondary markets (Serrano, 2010). The intuition again is straightforward. A patent is influential when it becomes a template that later inventors find it convenient to reuse, and it is broadly influential when that reuse crosses technology class boundaries rather than recycling within one niche.

Nineteenth century France lacks systematic citations and full abstracts, so the usual measures are unavailable. My strategy is to recover the same economic primitives using the language every patent supplies. Using patent titles, I build backward and forward measures that align with the originality and generality intuition. This strategy follows the historical-innovation literature that reconstructs novelty and diffusion when modern citation data are unavailable (MacLeod and Nuvolari, 2016; Nuvolari and Tartari, 2011).

My approach builds on the framework developed by Kelly et al. (2021), who use text-based backward and forward similarity to measure novelty and influence when citation data are not available.[7] Their core insight is simple. A patent is novel when its language differs from what came before it, and it is influential when later patents pick up its language. La Mela et al. (2024) apply this logic to historical Swedish patents.[8]

[7] Kelly et al. (2021) measure patent quality using the language patents share across time. They represent each patent's text as a weighted list of words, where unusual technical terms receive more weight than common ones. Critically, they weight terms according to how common they were *up to* the moment the patent was filed, so that early uses of new terminology are recognized as novel rather than penalized because the terms later become widespread. They then measure how similar two patents are by comparing these weighted word lists. For a patent filed in year t, they sum the similarity scores with all patents in the prior five years to get a backward similarity score, and sum similarities with all patents in the subsequent years to get a forward similarity score. Quality is the ratio of forward to backward similarity, adjusted by removing year-specific averages. They show this measure predicts future citations and correlates with market value. Their approach relies on adding up raw similarity scores, which means a few very similar patents can dominate the measure, and their year-adjustment makes cross-period comparisons difficult. My threshold-based counting approach with explicit exposure adjustment addresses both issues while preserving their core insight.

[8] La Mela et al. (2024) adapt the method of Kelly et al. (2021) to historical Swedish patents (1890–1929). They make one important modification: instead of summing similarity scores, they average them by dividing by the number

I follow the same intuition. Before laying out the full procedure in the next section, I briefly summarize how my implementation differs in five ways that matter for historical contexts and patent titles.[9] First, where they measure similarity by counting shared words, I compare the underlying meaning of titles. This allows me to see that "vulcanizing rubber" and "hardening caoutchouc" describe the same idea even when they share no words. Second, some titles use generic language that resembles many others, such as "improved method for manufacturing." I apply a correction that rewards pairs of patents that are specifically similar rather than generically similar.[10] Third, instead of adding up all similarity scores with neighbors, I count how many neighbors are truly close, close enough that the match is unlikely to occur by chance. This yields an interpretable quantity: the number of semantic ancestors or descendants a patent has. Fourth, I adjust these counts for opportunity. A patent filed during a busy period has more chances to resemble something than one filed in a quiet year, so each count is divided by the number of opportunities it faced. Fifth, I allow patents to belong partly to several technology classes at once, which better reflects multi-purpose inventions. These differences produce novelty and influence measures that remain comparable across time periods and technology fields.

3.2.1 Novelty

I construct a backward-looking measure of a patent's novelty that asks a simple question: at the moment a patent is filed, how much does its title echo ideas that were already present, and does that resemblance sit within the patent's own broad technology class or spill across different ones. To answer that, I translate patent titles into a common idea space, look back twenty years, and count the earlier titles that land unusually close to the new patent's title. I then scale those counts by the number of opportunities there were for resemblance, since some years and technology classes offer far richer prior art than others, and I finally express the result as a percentile within the patent's own year and technology class. A value at the 90th percentile means the patent's title resembles much of its accessible past. I define novelty as one minus this backward percentile, so higher values correspond to greater originality. Formally, I refer to the underlying resemblance process as Backward Semantic Adoption (BSA). I use BSA_g, BSA_{in}, and BSA_x for the global, within-class, and cross-class components.

The mechanics mirror familiar normalization problems in economics, namely separating signal from exposure and congestion. I begin with all patents' titles and their application years. Each patent carries a vector of memberships across twenty broad technology classes, denoted $\mathbf{f}_i = (f_{i1}, \ldots, f_{iC})$ with $\sum_c f_{ic} = 1$.[11] I do not treat class membership as all-or-nothing. Instead, a patent can be partly associated with several classes, with each element f_{ic} representing the strength of its connection to class c. When I refer to within-class resemblance, I mean links that fall in the same technology classes where the new patent is most strongly represented, while cross-class resemblance refers to links that occur in other classes. For percentile comparisons, I assign each patent a primary technology class, defined as the class with the highest membership share, purely to identify its peer group.

Patent titles are mapped into a high-dimensional semantic space using a language model I fine-tuned on patent text from the same historical period as my data.[12] Each title i is represented by a

of patents in each time window. This adjustment accounts for the fact that patenting activity grew substantially over their period. They validate their measures using patent renewal data and a list of famous Swedish inventors, finding that their importance measure correlates with how long patents were kept in force. However, their averaging does not account for shifts in the technology mix: a patent in a crowded technology class faces more potential connections than one in a sparse class. My opportunity denominators adjust for both the total number of available patents and their distribution across technology classes, producing measures that are robust to changes in both patenting intensity and the technological composition of innovation.

[9] This is also done more in line with the development of machine learning methods for Out-of-Distribution detection and Novelty Detection. See Barcina-Blanco et al. (2024) for a recent review.

[10] The correction is known as Cross-domain Similarity Local Scaling (CSLS) in the machine translation literature, where it was developed to match words across languages. In the context of my measure, a patent pair receives a high adjusted score only if they are close to each other but not close to many others. In practice, this correction substantially reduces false positives from generic phrasing. See Section 3.2.1.

[11] These class memberships use the twenty-sector technological classification defined by INPI and documented in Merouani and Perrin (2026). I train a patent–technology–class classifier on historical patent data to assign memberships to these sectors. The model achieves an F1 score of 0.95, and I interpret each element f_{ic} as the strength of patent i's connection to class c.

[12] I domain-adapted the InfoXLM-Large language model on all French and Italian patent titles filed up to 1930. This exposes the model to the multilingual technical vocabulary of the era and substantially improves its ability to

normalized embedding vector $\hat{\mathbf{e}}_i$, and closeness between two titles is measured by cosine similarity $s_{ij} = \hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_j$. Two titles are close if they describe the same idea in different words. For example, "process for vulcanizing rubber" and "method to harden caoutchouc" would sit near each other. For a patent filed in year t, I consider all earlier patents filed between $t-20$ and $t-1$ and down-weight those that appear further back by $w(\Delta) = 1/(1+\Delta)$, where $\Delta = t - y_j$. This gives more weight to near-term antecedents and less to distant ones. To preserve strict time use, all computations for a patent use only information available up to its filing year.

Nearest-neighbor methods in such spaces have a well-known bias: some titles act as hubs that look similar to many others simply because they contain common language. To correct this, I adjust each similarity by how generically close both titles are on average. The adjusted measure, known as the Cross-domain Similarity Local Scaling correction (CSLS), is

$$s_{ij}^{\mathrm{CSLS}} = 2s_{ij} - r_i - r_j,$$

where r_i is the new patent's average similarity to its nearest past neighbors and r_j is the past patent's average similarity to its nearest seeds in the same rolling window. This adjustment rewards pairs that are specifically similar rather than generally similar to everything. A pair like "process for vulcanizing rubber" and "method to harden caoutchouc" retains a high adjusted similarity, while a generic title such as "method for manufacturing" is penalized because it sits close to many unrelated patents. In practice, I compute r_i within each seed year using the full set of potential predecessors in the twenty-year window. The reverse averages for past patents, which would otherwise require recomputing similarities against all seeds, are estimated using a subsample of seeds within each batch. This approximation keeps the results unchanged in expectation while making the computation manageable.

I define an ancestral link as a case where the adjusted similarity s_{ij}^{CSLS} between a new patent and an earlier patent is unusually high for that time slice. Rather than choosing this cutoff arbitrarily, I determine it empirically so that only one percent of random pairs would exceed it by chance. For each filing year t, I take a large random sample of seed–past pairs within the twenty-year window, compute their adjusted similarities, and set the global threshold $\tau_g^B(t)$ at the 99th percentile of that empirical distribution. Because the typical degree of similarity varies across technology classes, I also estimate class-specific thresholds. For each class c, I compute a within-class threshold $\tau_{\mathrm{in}}^B(c)$ that gives greater weight to past patents more strongly associated with that same class. I further define cross-class thresholds $\tau_x^B(c_s, c_f)$ that depend on how strongly the seed patent belongs to one class c_s and the past patent to another c_f, using weights proportional to $f_{ic_s} f_{jc_f}$ with $c_s \neq c_f$. This weighting emphasizes strong cross-links between well-defined classes rather than faint overlaps. All thresholds are estimated separately for each seed year so that they adjust automatically to gradual shifts in language use and the distribution of technology classes over time.

With the thresholds established, I count how many earlier patents fall above them. For each patent i, I identify its 5,000 nearest predecessors in the semantic space and apply the same time-decay weights, giving more importance to links that occur closer in time. The global raw count

$$\mathrm{BSA}_g^{\mathrm{raw}}(i) = \sum_j w_{ij} \mathbf{1}[s_{ij}^{\mathrm{CSLS}} \geq \tau_g^B(t_i)]$$

adds up these time-weighted matches that exceed the global threshold. Each term w_{ij} discounts more distant years, and the indicator marks whether the adjusted similarity is high enough to qualify as an ancestral link.

To describe how resemblance sits within or crosses technological classes, I separate these links into two groups. The within-class raw count

$$\mathrm{BSA}_{\mathrm{in}}^{\mathrm{raw}}(i) = \sum_j \sum_c w_{ij} f_{ic} f_{jc} \mathbf{1}[s_{ij}^{\mathrm{CSLS}} \geq \tau_{\mathrm{in}}^B(c)]$$

captures reuse that remains inside the same broad areas. Each passing pair is weighted by the

identify semantic similarity across languages and historical terminology.

overlap of the seed's and the past patent's class profiles. The cross-class raw count

$$\mathrm{BSA}_x^{\mathrm{raw}}(i) = \sum_j \sum_{c_s \neq c_f} w_{ij} f_{ic_s} f_{jc_f} \mathbf{1}[s_{ij}^{\mathrm{CSLS}} \geq \tau_x^B(c_s, c_f)]$$

captures cases where the new patent resembles ideas from different areas, such as a machinery title echoing an earlier chemical process. Because I only evaluate the 5,000 closest predecessors for each patent, the resulting counts are conservative lower bounds on total resemblance. In practice, true links nearly always lie among those nearest neighbors, so the bound is tight.

Raw counts must be adjusted for exposure. A patent filed in a year that follows many earlier patents simply has more chances to resemble something than one that follows a quieter period. To correct for this, I divide each raw count by an opportunity denominator that reflects how much prior art existed in the relevant window, using the same time-decay weights. The global backward opportunity for patents filed in year t is

$$N_g^B(t) = \sum_{\Delta=1}^{H} w(\Delta) n_{t-\Delta}, \quad H = 20,$$

where $n_{t-\Delta}$ is the number of patents filed Δ years earlier. For within- and cross-class resemblance, I adjust for how much past "mass" exists in each class. The within-class opportunity for patent i is

$$N_{\mathrm{in}}^B(i) = \mathbf{f}_i^\top \mathbf{N}_{\mathrm{class}}^B(t_i), \quad \text{where} \quad \mathbf{N}_{\mathrm{class}}^B(t) = \sum_{\Delta=1}^{H} w(\Delta) \sum_{j: y_j = t-\Delta} \mathbf{f}_j,$$

and the cross-class opportunity is the remainder once within-class opportunities are subtracted:

$$N_x^B(i) = \Big(\sum_c N_{\mathrm{class},c}^B(t_i) \Big) - N_{\mathrm{in}}^B(i).$$

These denominators capture how many opportunities existed for a patent's title to echo the past, both within its own technological areas and across others.

I then compute exposure-adjusted backward rates as

$$\mathrm{BSA}_\ell^{\mathrm{rate}}(i) = \frac{\mathrm{BSA}_\ell^{\mathrm{raw}}(i)}{N_\ell^B(i)}, \quad \ell \in g, \mathrm{in}, x.$$

This turns each raw count into a rate that measures the share of available prior art that the patent closely resembles.

Finally, I express these rates as percentiles within each patent's own year and main technology class. A patent with a global backward rate percentile of 0.90 resembles much of what was already available to it, and is therefore relatively unoriginal. To make interpretation intuitive, I define novelty as one minus this percentile,

$$\mathrm{Novelty}(i) = 1 - \mathrm{BSA}_g^{\mathrm{rate\_pct}}(i)$$

A value of 0.90 thus indicates that the patent ranks among the top ten percent of its contemporaries in linguistic originality. The resulting novelty score lies on a common 0 to 1 scale and is directly comparable across filing years and primary technology classes.

3.2.2 Influence

I measure a patent's influence by asking the forward question of the novelty measure. Instead of asking how much a patent draws on ideas that already existed, I ask how much later inventions draw on it. The goal is to quantify how often a patent's language is taken up by its successors, and whether that take-up remains within the patent's broad technology class or moves across others. As before, I represent titles in a common idea space and look forward twenty years to identify later titles that land unusually close to the seed patent's title. These forward matches, weighted by their temporal proximity, form the basic indicator of how much a patent's phrasing reappears in the

subsequent record. I refer to this process as Forward Semantic Adoption (FSA) and I focus on the global component as the headline measure.

The computation follows the same steps as in Novelty, except that time now runs forward. For a patent i filed in year t_i, I compare its embedding to those of patents j filed between $t_i + 1$ and $t_i + 20$.[13] Similarities are adjusted with the same CSLS correction so that generic phrasing does not inflate forward closeness, and I apply the same time-decay weight,

$$w_{ij} = \frac{1}{1 + (y_j - t_i)}.$$

I then define the forward adoption counts using the same empirical thresholds that were used for Novelty, now sampled from the future of each filing year. The global headline count is time-weighted and thresholded:

$$\mathrm{FSA}_g^{\mathrm{raw}}(i) = \sum_j w_{ij}\, \mathbf{1}\left[s_{ij}^{\mathrm{CSLS}} \geq \tau_g(t_i)\right],$$

with within-class and cross-class analogues constructed in the same way if desired. Because I evaluate only the nearest neighbors, these are conservative lower bounds that are tight in practice.

To separate traction in the future from inherited language at birth, I scale the forward signal by the patent's own backward resemblance from the novelty section. Let $\mathrm{BSA}_g^{\mathrm{raw}}(i)$ be the global backward count defined there. The baseline influence signal is the log ratio

$$I(i) = \log\Big(1 + \frac{\mathrm{FSA}_g^{\mathrm{raw}}(i)}{\mathrm{BSA}_g^{\mathrm{raw}}(i) + 1}\Big),$$

which captures how often a patent's ideas are reused per unit of inherited language. The +1 in the denominator avoids division by zero and dampens cases with very small backward counts.

I then place $I(i)$ on a common scale within broad time–class cells. Let $d(i)$ denote the decade of t_i and let $c(i)$ be the patent's primary technology class. Within each (d, c) cell I compute a robust z-score using the median and the median absolute deviation (MAD),

$$\mathrm{Influence}(i) = \frac{I(i) - \mathrm{median}_{d,c}[I]}{1.4826 \cdot \mathrm{MAD}_{d,c}[I] + \varepsilon},$$

where 1.4826 rescales the MAD to the standard deviation under normal tails and ε is a small constant for numerical stability.[14]

Interpretation is direct. Positive values indicate that the patent's language was taken up more than is typical for patents from the same decade and primary class, conditional on how much it resembled the past when it arrived. Negative values indicate below-typical propagation. Because the statistic is standardized within decade and technology class, it is comparable across periods and technology classes without further adjustments. Within-class and cross-class versions can be reported by replacing $\mathrm{FSA}_g^{\mathrm{raw}}$ with the corresponding forward counts in the construction above.

3.2.3 Do the measures hold?

A credible influence measure should rank the inventions that historians already treat as landmarks above ordinary patents, for example Pasteur's preservation method, the Jacquard programmable loom, or the Lumière cinématographe. I benchmark the influence score against forty-one notable nineteenth-century patents drawn from standard histories and encyclopedias of invention.[15] The set spans chemistry, metallurgy, precision instruments, machinery, navigation, construction, lighting and heating, and road transport. None of these labels enter the construction of the measure; they serve only as an external yardstick for whether the language of influential inventions echoes forward

[13] I avoid issues of right censoring by expanding the patent database on the level of patent titles through the 1930s.

[14] The robust standardization uses $\mathrm{MAD} = \mathrm{median}(|I - \mathrm{median}(I)|)$. The factor 1.4826 puts the MAD on the same scale as a standard deviation under a Gaussian benchmark. Numerically, a value of +1 on the robust z-scale is about one robust standard deviation above the median, which corresponds to roughly 1.4826 MAD above the median. I set $\varepsilon = 10^{-12}$.

[15] Table 15 in Appendix D lists the forty-one benchmark patents, their short titles, technology classes, and application codes.

into later titles. If the score captures technological reach, known landmarks should float to the top without being hard-coded into the measure.

I summarize performance with rank-based accuracy. The area under the ROC curve (AUC) is the probability that a randomly chosen famous patent receives a higher influence score than a randomly chosen non-famous patent.[16] Intuitively, draw one famous and one ordinary patent at random; AUC is the chance the score puts the famous one higher. Across the full 1790 to 1899 sample, the influence score attains an AUC of 0.609. Focusing on the period where forward semantic adoption is most informative, 1850 to 1899, the pooled (micro) AUC rises to 0.673, and averaging by decade in that window yields a macro AUC of 0.693.[17] A permutation check that shuffles the fame labels within the late window centers near 0.500 with a dispersion of about 0.056, which confirms that the observed separation is not a mechanical artifact of sampling. Permutation functions as swapping name tags on the same patents; if the score had no signal, its rank accuracy would match the shuffled baseline. I focus on 1850 to 1899 for headline results because the twenty-year forward window is well populated; results are similar when I restrict to patents with a complete forward window of twenty years.

The signal strengthens over time. By decade, late-century accuracies are highest, with an AUC of 0.872 in the 1880s, and solid performance earlier in the window, for example 0.702 in the 1850s and 0.660 in the 1860s. By technology class, the measure performs best where imitation and adaptation were active. Marine and navigation, mining and metallurgy, chemical arts, and precision instruments all show high rank accuracy. Where ideas were traded and reused more often, language leaves more footprints, and a semantic measure can see those traces. In classes with only a single famous case, such as agriculture or construction in this sample, AUCs are unstable and should be interpreted cautiously. Low AUC in a cell with one famous case reflects small-n volatility, not a change in construct validity.

I next ask whether the influence score adds information beyond fixed effects and a control for how derivative the patent was at birth. I estimate a logistic model for fame with decade and major-class indicators and log backward resemblance as baseline predictors. This asks whether the score helps once we already know when and where a patent was filed and how derivative its title was. In five-fold cross-validation over 1850 to 1899 the baseline attains an AUC of 0.629. Adding the standardized influence score raises the cross-validated AUC to 0.661, an improvement of 0.032. At this level of class imbalance, a three-point AUC gain is economically meaningful. Out-of-time tests tell the same story. Training the model on filings through 1879 and evaluating on subsequent decades yields an AUC of 0.615 with influence included versus 0.571 without it. Training through 1889 and testing after yields 0.579 with influence versus 0.514 without. Training on earlier decades and testing later asks whether a rule learned before widespread electrification and large-scale steel production still flags the right inventions afterward. A cluster bootstrap that resamples decade-by-class cells produces an average AUC of 0.612 with a 95 percent interval from 0.519 to 0.708, which indicates that the separation is not driven by a handful of clusters.[18] Resampling whole decade-by-class blocks preserves within-cluster correlation, so the interval reflects genuine uncertainty rather than overconfident independence. The construction is stable across reasonable choices of the ridge parameter in the denominator and of the winsorization cutoffs. Tuning within 1850 to 1889 selects $\lambda = 1$ and delivers the same headline AUCs.

Tail behavior is consistent with these rank statistics. In the late window, famous patents are several times more frequent in the upper tail of the score distribution than in the population as a whole. In pooled enrichment curves, the top one to two percent of patents by the influence score contain famous inventions at roughly six to seven times the base rate. A lift above one means famous patents are more common in that bin than in the population; for example, lift of three triples the base rate. Figure 2 shows the decile lift for 1850 to 1899, with a monotone decline across deciles,

[16] An AUC of 0.50 corresponds to random ranking. An AUC of 0.67 means the score correctly orders about two-thirds of random famous versus non-famous pairs. Under class imbalance, AUC remains a stable pairwise-ranking metric, which is why I foreground it and use lift and gains curves to summarize the tail.

[17] The pooled (micro) AUC treats every patent equally across decades. The macro AUC averages decade-specific AUCs, giving each decade equal weight.

[18] I re-estimate the score using $\lambda \in 0.5, 1, 2, 5$ in the ratio $\log(1 + \frac{\text{FSA}}{\text{BSA}+\lambda})$. For winsorization of the log-ratio, I compare percentiles $(0.1, 99.9)$ and $(1, 99.5)$. Headline AUCs for 1850–1899 are identical to three decimals across these settings, and the by-decade ordering is unchanged. Ridge λ controls variance when BSA is near zero; winsorization protects the robust z from extreme ratios. Both are standard regularization choices.

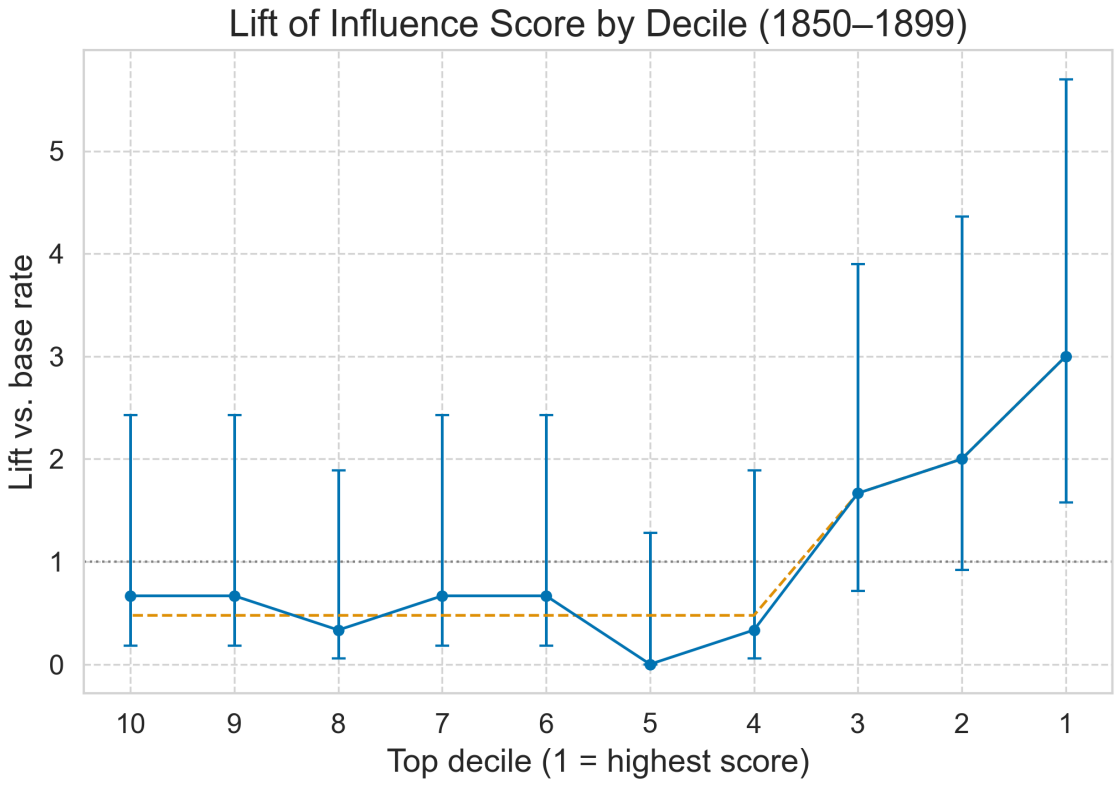


Figure 2: Concentration of Historically Famous Patents by Influence Decile

a top decile well above one, and Wilson confidence intervals that do not overlap the base rate.[19] At extremely small K, precision at K is mechanically low because famous cases are rare. The absence of hits at K in per-decade 0.1 percent slices is therefore not informative by itself; the pooled enrichment and gains curves are the correct diagnostics under this class imbalance.[20]

As a robustness check, I re-express the forward signal as a residual. Intuitively, this isolates the part of forward take-up that is not explained by how derivative the language was or by time and technology class. I regress log forward counts on log backward counts with decade and class fixed effects, take residuals, and standardize within decade by class. This alternative "excess influence" statistic achieves an AUC of 0.668 in the late window, which is nearly identical to the primary score, and confirms that the results are not specific to any one transformation.

The influence score behaves in line with its intent. It assigns higher values to inventions that historians have judged to be important, it performs best in periods and technology classes where diffusion was active, it adds information beyond fixed effects and baseline dependence, and it remains stable in out-of-time checks. The signal is moderate rather than perfect, which is appropriate given the small benchmark set and the extreme rarity of famous cases. Within those limits, the measure tracks the forward reach of ideas as they appear in language and move through the subsequent record.

4 What Collaboration Changes: Switching from Solo to Team

I begin by exploring what happens when inventors switch from working alone to working with others. In this study, working with others is measured as co-patenting. sI focus on inventors who filed at least one solo patent before their first collaboration and track four outcomes: annual patent output,

[19]The plotted intervals use Wilson binomial confidence intervals for per-decile hit rates, and the dashed curve is an isotonic smooth of lift across deciles.

[20]Famous patents are extremely rare relative to the population. Under such imbalance, precision-recall summaries and pooled lift or gains curves are more informative than hit counts at very small K within tiny strata.

the novelty and influence of their work, and whether they move outside their home technology class. The sample runs from 1791 to 1900 and includes inventors with at least two patents, the minimum needed to define a career span and estimate within-inventor effects.[21]

Throughout this section, I use event-study designs that align inventors by the year of their first collaborative patent, T_i. Event time $k = t - T_i$ measures years before or after this switch, with $k = 0$ marking the year of first collaboration. I estimate Sun–Abraham (SA) stacked specifications that compare early switchers to inventors who switch later, controlling for inventor fixed effects, either calendar-year or pre-event technology-class-by-year fixed effects, and stack (cohort-comparison) fixed effects. Standard errors cluster at the inventor level. The SA design accounts for staggered timing: different inventors start collaborating in different years, and I stack cohort-specific comparisons to form a single estimate (Sun and Abraham, 2021).

4.1 No Sustained Productivity Gain

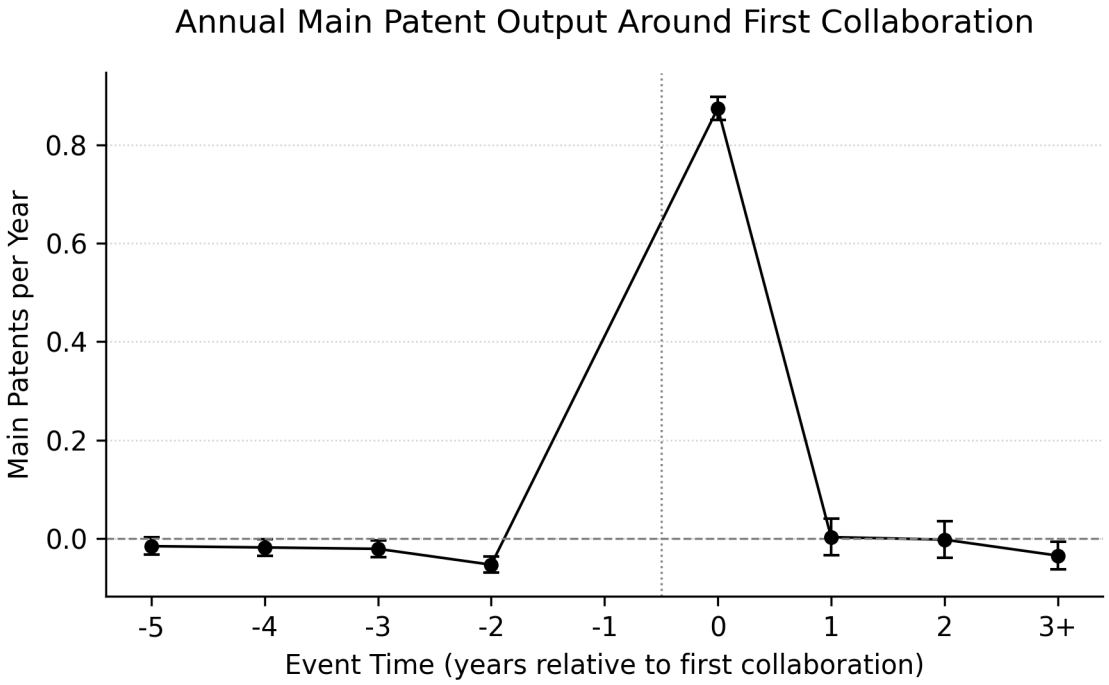


Figure 3: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is the number of main patents filed by inventor i in year t. Event time $k = 0$ marks the year of first collaboration; $k = -1$ is the omitted category. The sample includes inventors who filed at least one solo patent before collaborating (switchers) and inventors who never collaborate (never-treated controls), 1791–1900. The specification absorbs inventor fixed effects and calendar-year fixed effects. Event times $k \geq 3$ are binned into a single coefficient. Standard errors are clustered at the inventor level. The pre-event mean ($k \leq -2$) is 0.27 patents per year. Pre-trends are rejected ($F = 15.3$, $p < 0.001$). $N = 3,568,668$ inventor-year observations.

Figure 3 plots annual main-patent output around the first collaboration. The outcome is MainCount_{it}, the number of main patents inventor i files in year t. The estimating equation is:

$$y_{it} = \sum_{k \in \{-5,-4,-3,-2,0,1,2\}} \beta_k D_{it}^k \; + \; \beta_{3+} D_{it}^{k \geq 3} \; + \; \alpha_i + \gamma_t + \delta_s + \varepsilon_{it},$$

[21] Sample sizes differ across specifications because of data and design. All four outcomes (productivity, novelty, influence, and diversification) use main patents only. The productivity sample is largest (N = 3,568,668 inventor-years) because years with no main patents yield a valid zero for MainCount_{it}. Influence and novelty use smaller samples (N = 1,057,062 and 1,057,035 respectively) because years with no main patents produce missing values, one cannot compute a mean influence or novelty score when there are no patents, which the regressions drop. Diversification is smallest (N = 98,052) because it excludes never-treated inventors and requires non-missing pre-modal class assignment, which depends on having filed at least one patent before $k \geq -1$. These differences reflect outcome definitions and modeling choices, not selective data loss.

where D_{it}^k equals one if inventor i is k years from first collaboration and zero otherwise, α_i captures inventor fixed effects, γ_t captures calendar-year fixed effects, δ_s captures stack (cohort-comparison) fixed effects, and $k = -1$ is omitted. I use year fixed effects (rather than technology-class-by-year) because patent count is not mechanically related to technology-class choice: a patent contributes to the count regardless of which class it falls in. Year fixed effects control for aggregate trends, for example, nationwide surges in patenting activity, without conditioning on post-treatment class decisions. I bin event times at or beyond $k = 3$ into a single coefficient to preserve precision.

The pattern is clear. Before collaboration, switchers average 0.27 main patents per year. At $k = 0$, output spikes to $\hat{\beta}_0 = 0.87$ (s.e. 0.01). This is largely explained by the first collaborative patent itself, which accounts for most of the jump. After $k = 0$, the effect disappears. At $k = 1$, the coefficient is $\hat{\beta}_1 = 0.00$ (s.e. 0.02); at $k = 2$, it is $\hat{\beta}_2 = -0.00$ (s.e. 0.02); and for $k \geq 3$, it is $\hat{\beta}_{3+} = -0.03$ (s.e. 0.01). None of these post-event coefficients differs meaningfully from zero.

The specification fails the parallel-trends test. A joint test that pre-event coefficients equal zero rejects strongly ($F = 15.3$, $p < 0.001$). The negative pre-event coefficients suggest that inventors about to collaborate were already on a different trajectory from those who would switch later—perhaps slowing down to prepare a joint project. Even aside from the pre-trend failure, excluding the first collaborative patent removes the spike, confirming the $k = 0$ jump is mechanical. The first collaborative patent shows up in the count.

Two robustness checks confirm this interpretation. First, excluding the first collaborative patent from the count removes the spike entirely (Appendix E.1, Figure 12). Second, top-coding output at three patents per year leaves the event-year effect essentially unchanged ($\hat{\beta}_0 = 0.85$, s.e. 0.01), but the pre-trend problem remains (Appendix E.1, Figure 13). A more conservative specification that uses only switchers and absorbs inventor and technology-class-by-year fixed effects fails the pre-trend test and delivers a much smaller event-year effect ($\hat{\beta}_0 = 0.08$, s.e. 0.01; Appendix E.1, Figure 14).

The conclusion is straightforward: collaboration does not durably raise measured productivity. The apparent jump at $k = 0$ reflects the collaborative patent itself, not a permanent increase in output.

5 No Increase in Novelty

If collaboration leads to more original ideas, we might expect patents filed during or after the first team project to be more novel. That is, more distinct from prior art. I test this using a patent-level novelty score that measures how much a patent's language and concepts differ from the existing corpus at the time of filing. I construct it as the inverse percentile rank of backward semantic adoption: patents that draw less heavily on prior language score higher. For each inventor-year, I average novelty across all main patents filed that year.

I estimate a Sun–Abraham stacked specification using switchers and never-treated controls from 1791 to 1900. I absorb inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects. I absorb inventor FE and pre-event-modal-class-by-year FE to avoid conditioning on post-treatment class choice. Event times $k \geq 3$ are binned into a single coefficient, and standard errors are clustered by inventor.

Figure 4 shows essentially no effect. Coefficients at all event times are small and statistically indistinguishable from zero. At $k = 0$, the estimate is $\hat{\beta}_0 = -0.01$ (s.e. 0.01); at $k = 1$, it is $\hat{\beta}_1 = -0.00$ (s.e. 0.01); at $k = 2$, it is $\hat{\beta}_2 = 0.01$ (s.e. 0.01); and for $k \geq 3$, it is $\hat{\beta}_{3+} = -0.01$ (s.e. 0.01). Pre-event coefficients are similarly small and mixed in sign, with no evidence of pre-trends. The pattern holds when I split novelty into within-class and cross-class components (Appendix E.2, Figures C1–C2).

This null result is important. It rules out an originality-improvement story: collaboration does not make inventors' work systematically more novel or original.

5.1 An Immediate but Temporary Rise in Influence

Influence measures how much future inventors draw on a patent's ideas. I construct it as a z-score: for each patent, I compute the log ratio of forward semantic adoption (how much future patents

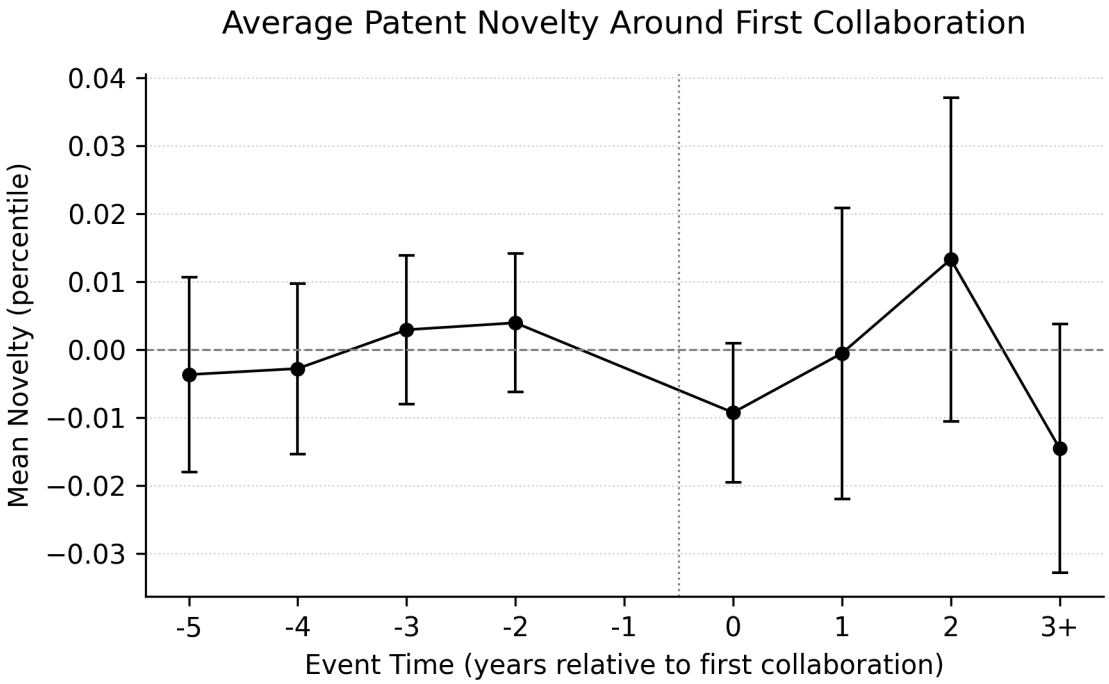


Figure 4: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is the mean novelty score across all main patents filed by inventor i in year t. Novelty is a patent-level percentile rank (0 to 1) measuring how distinct a patent's language and concepts are from prior art, constructed as the inverse of backward semantic adoption: higher values indicate greater originality. Event time k=0 marks first collaboration; k=-1 is omitted. The sample includes switchers and never-treated controls, 1791–1900. The specification absorbs inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects to avoid conditioning on post-treatment class choice. Event times k 3 are binned into a single coefficient. Standard errors are clustered at the inventor level. Coefficients are small and statistically indistinguishable from zero at all event times, indicating no systematic change in patent originality following collaboration. N=1,057,035 inventor-year observations.

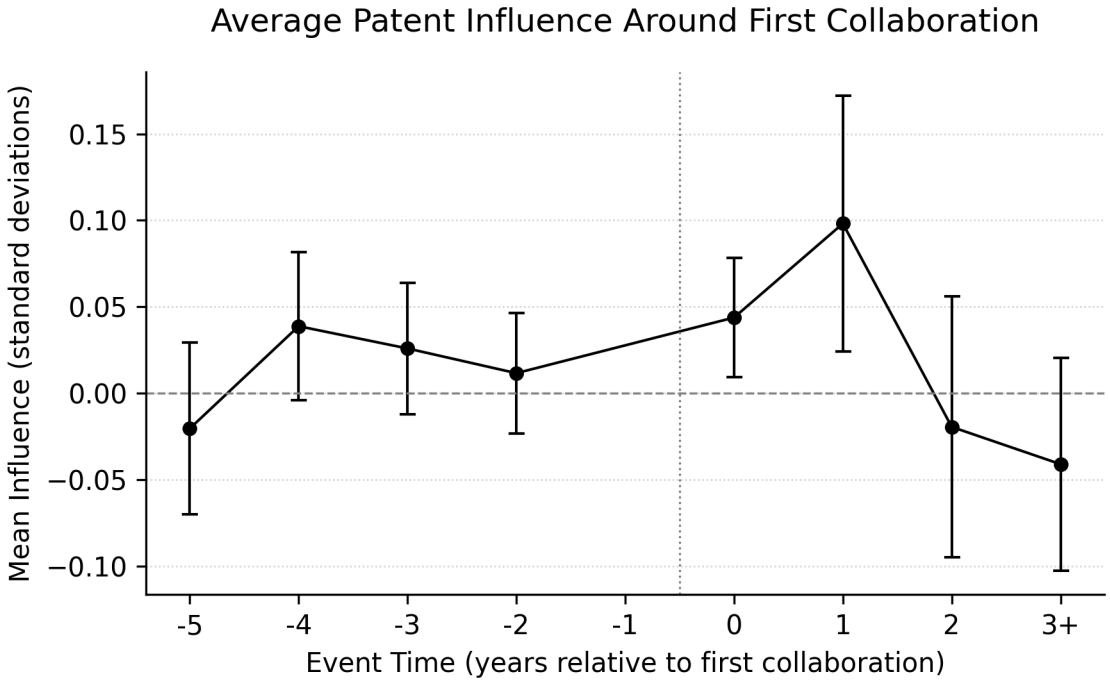


Figure 5: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is the mean influence score across all main patents filed by inventor i in year t. Influence is a patent-level z-score (standardized within decade and technology class) measuring how intensively future patents draw on a patent's ideas, constructed as the log ratio of forward to backward semantic adoption. Event time $k = 0$ marks first collaboration; $k = -1$ is omitted. The sample includes switchers and never-treated controls, 1791–1900. The specification absorbs inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects to avoid conditioning on post-treatment class choice. Event times $k \geq 3$ are binned. Standard errors are clustered at the inventor level. $N = 1,057,062$ inventor-year observations.

borrow from this patent's language) to backward semantic adoption (how much this patent borrows from prior patents' language), then standardize within decade and technology class. Higher values mean greater downstream impact.

For each inventor-year, I average influence across all main patents filed that year. I then estimate a Sun-Abraham stacked event study. I absorb inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects. I use this specification, rather than current-technology-class-by-year, to avoid conditioning on post-treatment class choice, which the diversification analysis below demonstrates is affected by collaboration. The pre-event modal class is the technology class where the inventor filed the most main patents before first collaboration ($k \leq -1$). Using pre-event-modal-class-by-year fixed effects controls for technology-specific time trends (e.g., surging patent activity in mechanical engineering during the 1860s) while avoiding "bad controls": absorbing the inventor's current class would condition on an outcome of collaboration itself. Event times $k \geq 3$ are binned, and standard errors are clustered by inventor.

Figure 5 shows the path. At $k = 0$, influence rises by $\hat{\beta}_0 = 0.04$ standard deviations (s.e. 0.02). It peaks at $k = 1$ with $\hat{\beta}_1 = 0.10$ (s.e. 0.04), then falls back: $\hat{\beta}_2 = -0.02$ (s.e. 0.04) is indistinguishable from zero, and estimates for $k \geq 3$ are small and imprecise. Pre-event coefficients are mixed and near zero, with no clear pre-trend.

The effect is immediate, but modest and short-lived. A placebo test that randomly assigns "first-collaboration" years to inventors who never collaborate produces a flat path, confirming the pattern is not spurious (Appendix E.3, Figure 17). A balanced-window specification that requires each switcher to have at least one solo patent before $k = 0$ and one collaborative patent after delivers nearly identical results (Appendix E.3, Figure 18).

In sum, collaboration brings an immediate, small, and temporary bump in influence, roughly 0.1 standard deviations at its peak, but no sustained step-change. This modest effect contrasts sharply with the large, persistent diversification across technology classes I document next.

5.2 Collaboration Diversifies Work Across Technology Classes

To see where collaboration redirects effort, I examine whether inventors move outside their home technology class. For each inventor, I define c_i^{pre} as their most frequent broad technology class (one of twenty categories) in the years before first collaboration. I then compute:

$$\text{OutsideClassShare}_{it} = \frac{\text{Number of main patents in year } t \text{ outside } c_i^{\text{pre}}}{\text{Total main patents in year } t}.$$

This share ranges from zero (all patents in technology home class) to one (all patents outside technology home class). Anchoring to c_i^{pre}, determined before collaboration, avoids conditioning on post-treatment choices and makes effects interpretable as percentage-point changes. Years with zero main patents have an undefined share and are excluded from this outcome.

The outcome is a share, bounded between zero and one. Unlike productivity (a count) or influence (a z-score), which can vary freely, shares may drift smoothly with career age: as inventors gain experience, they may naturally explore outside their home class, producing a gradual upward trend unrelated to collaboration. This secular drift could mask, or be mistaken for, a sharp collaboration-related jump. To isolate the collaboration effect, I first remove within-inventor linear time trends: for each inventor, I regress $\text{OutsideClassShare}_{it}$ on calendar year, then use the residuals as the outcome. This detrending ensures that coefficients capture deviations from each inventor's smooth trajectory, the sudden "kink" at collaboration, not general career-age exploration. I do not detrend productivity or influence because those outcomes are unbounded and do not mechanically trend in one direction with age.

I estimate a Sun–Abraham stacked event study using only not-yet-treated switchers as controls. I exclude never-collaborators here because they may differ fundamentally from switchers in ways that make them a poor counterfactual. Never-collaborators may be solo by choice, for example, working on highly specialized problems ill-suited to teamwork, possessing personality traits that discourage collaboration, or operating in geographic or social isolation. Or by constraint, such as lacking access to inventor networks. These factors likely affect their technology-class choices independently of

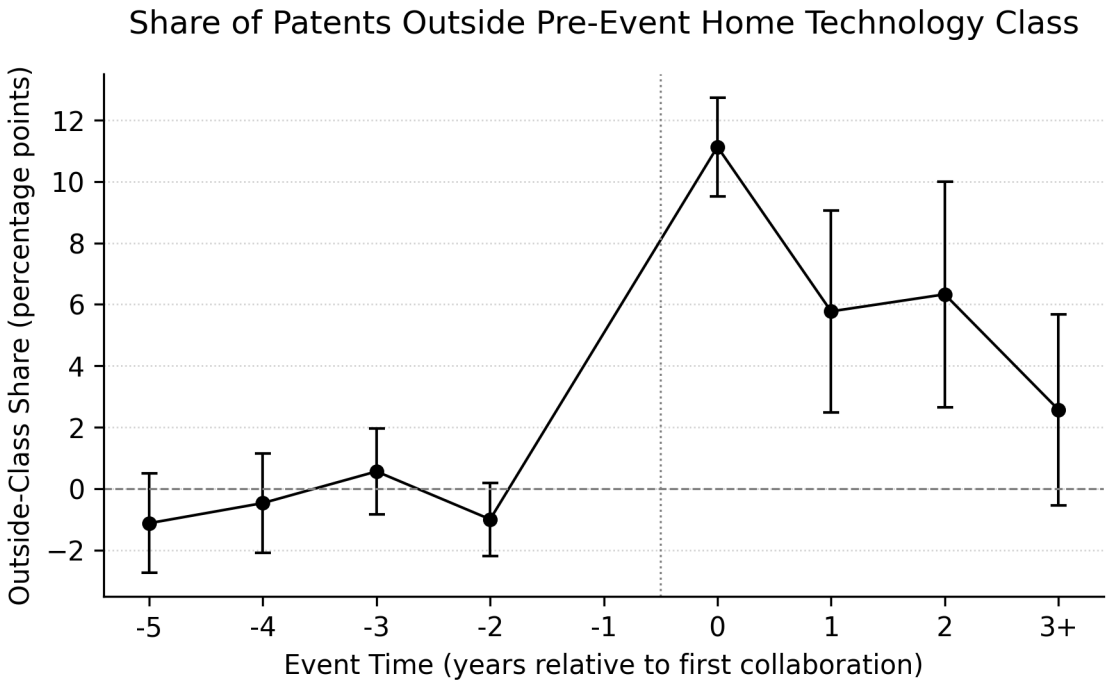


Figure 6: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is the share of inventor i's main patents in year t that fall outside their pre-event modal technology class, where the modal class is computed from all main patents filed in $k \leq -1$. The outcome is linearly detrended within inventor (by calendar year) before estimation to remove smooth career-age trends. Event time $k = 0$ marks first collaboration; $k = -1$ is omitted. The sample includes switchers only, 1791–1900, with not-yet-treated switchers as the sole control group (never-treated excluded). The specification absorbs inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects. Event times $k \geq 3$ are binned. Standard errors are clustered at the inventor level. Pre-trends are not rejected ($F = 1.5$, $p = 0.21$). $N = 98,052$ inventor-year observations.

any collaboration "effect." In contrast, not-yet-treated switchers are observably similar to treated switchers, they also eventually collaborate, probably differing mostly in the timing of their switch. This makes them a cleaner comparison group: they most likely show what switchers' outcomes would have looked like absent collaboration, holding constant the types of inventors who ultimately team up. I absorb inventor fixed effects and pre-event-modal-class-by-year fixed effects, bin $k \geq 3$, and cluster standard errors by inventor.

Figure 6 shows a sharp, persistent shift. At $k = 0$, the outside-class share jumps by 11.1 percentage points ($\hat{\beta}_0 = 0.111$, s.e. 0.01). The effect remains large at $k = 1$ (5.8 pp, s.e. 0.02) and $k = 2$ (6.3 pp, s.e. 0.02), then moderates slightly at $k \geq 3$ (2.6 pp, s.e. 0.02). A joint test on pre-event coefficients does not reject equality with zero ($F = 1.5$, $p = 0.21$).

This is economically meaningful. Before collaboration, switchers do some work outside their home class—the baseline is not zero. Collaboration roughly doubles or triples that share immediately and keeps it elevated for years.

Two diagnostics confirm robustness. First, a balanced-window specification that requires observable solo and collaborative patents on both sides of $k = 0$ delivers nearly identical results: 12.1 pp at $k = 0$, 6.0 pp at $k = 1$, 5.1 pp at $k = 2$, with no pre-trend problem ($F = 1.5$, $p = 0.20$; Appendix E.4, Figure 19). Second, a placebo that randomly shuffles first-collaboration years among switchers produces a flat path with a tiny, insignificant $k = 0$ blip of -0.2 pp (s.e. 2.8) and clean pre-trends ($F = 1.9$, $p = 0.11$; Appendix E.4, Figure 20). The diversification result is not an artifact of specification or sample composition.

The team composition aligns with this finding. When inventors first collaborate, they work with people from different technology classes. In parallel event studies, I find that the share of teammates whose prior work lies outside the focal inventor's home class rises by 1.4 percentage points at $k = 0$, the number of distinct prior technology classes represented on the team increases by about 0.01, and a normalized team-breadth index rises by roughly 0.01 at $k = 0$, with positive effects continuing at $k = 1$. Cross-class teammates predict cross-class output.

The bottom line is clear and robust across multiple specifications: collaboration does not make inventors permanently more productive (the $k = 0$ spike is mechanical), nor does it raise the novelty of their work (coefficients are flat and near zero). It brings a small, temporary influence bump that dissipates within two years. What collaboration does is diversify where inventors work. Switchers shift 10–12 percentage points of their output outside their historical home class immediately upon teaming up, and this redirection sustains for years. This diversification result survives balanced-window restrictions, placebo tests with permuted timing, and alternative control groups, and aligns with direct evidence that switchers team up with people from different technology classes. Diversification, not productivity or originality gains, is the primary channel through which collaboration changes inventive work. The next section examines the network structure that enables this cross-class collaboration.

6 Inventor's network

The event study shows that moving from solo work to collaboration leads inventors to spread their efforts across more technology classes and produce more influential early post-event patents, without generating persistent novelty gains. Collaboration also creates new ties. Once inventors start working with others, they build networks. Studying these networks is a natural next step for understanding the mechanisms behind the patterns in the event study.

I begin with the network architecture. The density, fragmentation, and degree distribution, before turning to who collaborates with whom, especially across gender and family lines. I then link these features to the diversification patterns documented in the event study.

6.1 Node to node: Structure of the Network

I construct an inventor network where two inventors are linked if they appear together as applicants on the same patent. The structure of the inventor network reveals that collaboration in 19th-century France was rare, fragmented, and highly unequal. Of the 254,547 inventors in the sample, 175,794

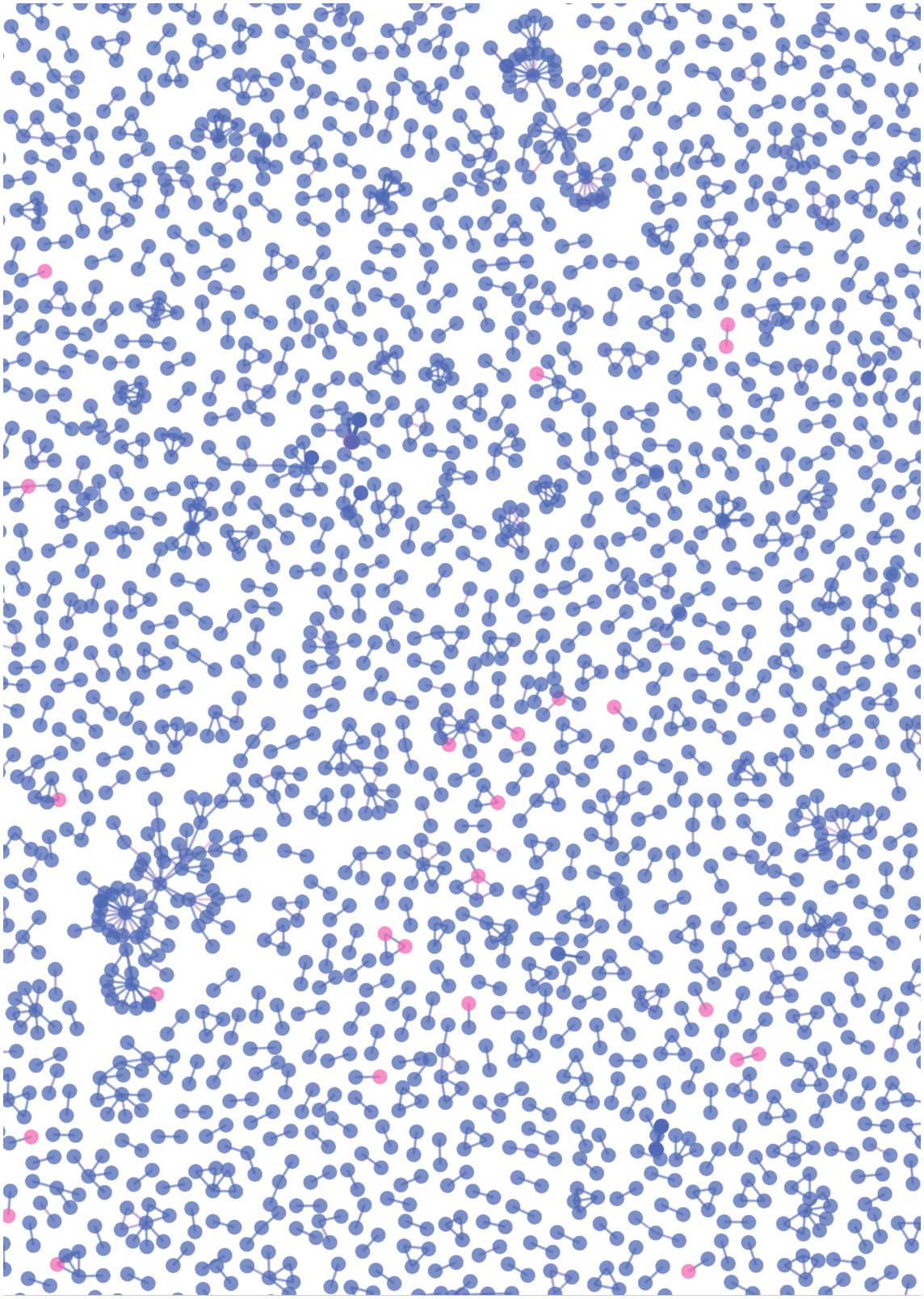


Figure 7: Visualization of the largest component of the network, shown on the left. Red nodes are women, blue nodes are men.

worked entirely alone, corresponding to 69.1% of all inventors. They filed their patents solo and improved their designs solo. The median inventor had zero collaborators.

Collaboration was not only rare but also limited in scope. Among those who did collaborate, there are 52,146 ties in total. This yields a network density of only 1.61×10^{-6}, meaning that almost none of the potential connections among inventors were realized. The average degree is just 0.41, so even within the collaborative minority, most inventors had only one or two co-inventors. The distribution of ties was also highly unequal: the 90th percentile inventor had one collaborator, and at the 99th percentile the number was just three collaborators.

The network was also highly fragmented. Inventors are split into 209,470 separate components. Most of these were minimal: 83.9% were isolates, 13.0% were dyads, and only 0.04% contained more than ten inventors. The median non-isolate component contained just 2 inventors. The largest component had 76 inventors. Within that group, however, ideas could travel quickly, since the average distance between any two inventors was only 4.09 steps.

Where collaboration did occur, it was concentrated in small, tightly knit teams. The average clustering coefficient is 0.06, which reflects that most inventors worked alone or with only one collaborator, leaving little room for their collaborators to be connected to each other. Global transitivity, however, is much higher at 0.64. This captures the fact that when patents did involve three or more inventors, these groups almost always formed closed triangles in which everyone collaborated with everyone else. Together, these measures show that teamwork was uncommon, but the teams that did exist were locally dense.

Finally, collaboration was not random with respect to inventors' connectedness. The degree assortativity of 0.37 indicates a positive correlation: inventors with many collaborators were more likely to work with others who also had many collaborators, rather than with isolated inventors.

Collaboration in 19th-century France was therefore both rare and fragmented. Why did so many inventors choose to work alone? And when they did collaborate, what determined their choice of partners, whether by gender, by technology class, or by other dimensions?

6.2 Who collaborates? Gender, family and technology

Conditional on collaborating, women built denser networks than men. Figure 8 shows that after the 1840s, women's average degree at times surpassed men's, and after the 1860s they did so more regularly. This occurred despite their rarity: women made up 1.7% of inventors and appeared on 1.8% of patents. Women were also more likely to work alone: 76.5% of women were isolates, compared to 68.9% of men.

Women's collaboration intensity was strongest in the very technology classes where they were least likely to appear. Figure 9 shows that in railways, machinery, marine & navigation, construction, and mining & metallurgy, women collaborated more intensively than men. Yet the findings of (Merouani and Perrin, 2026) confirm that these same technology classes were significantly less likely to have female-linked patents overall. Ceramics stands out as the exception: women collaborated more in this technology class, and their overall participation was neither higher nor lower than average. Together, these patterns suggest that women often used collaboration as a strategy to break into male-dominated technology classes.

Family ties provided the key entry point into inventive collaboration for many women. I mark edges as family ties if the connected inventors shared a surname, including maiden names. Just 9.5% of all ties were family links, but for women the share was 35.4%. For men it was only 8.9%. Families were an important channel into collaboration for everyone, but more so for women.

Collaboration was overwhelmingly between men, and cross-gender for women. Of all ties, 97.7% linked two men; only 1.9% were mixed gender and 0.37% linked two women. Among ties that included a woman, 84.0% connected her to a man. The modest gender assortativity of 0.266 reflects scarcity, not preference.

Collaboration in nineteenth-century France was profoundly gendered. Women inventors were scarce and often worked alone, but those who collaborated sometimes built networks as dense as men's, particularly in male-dominated technology classes and through family ties.

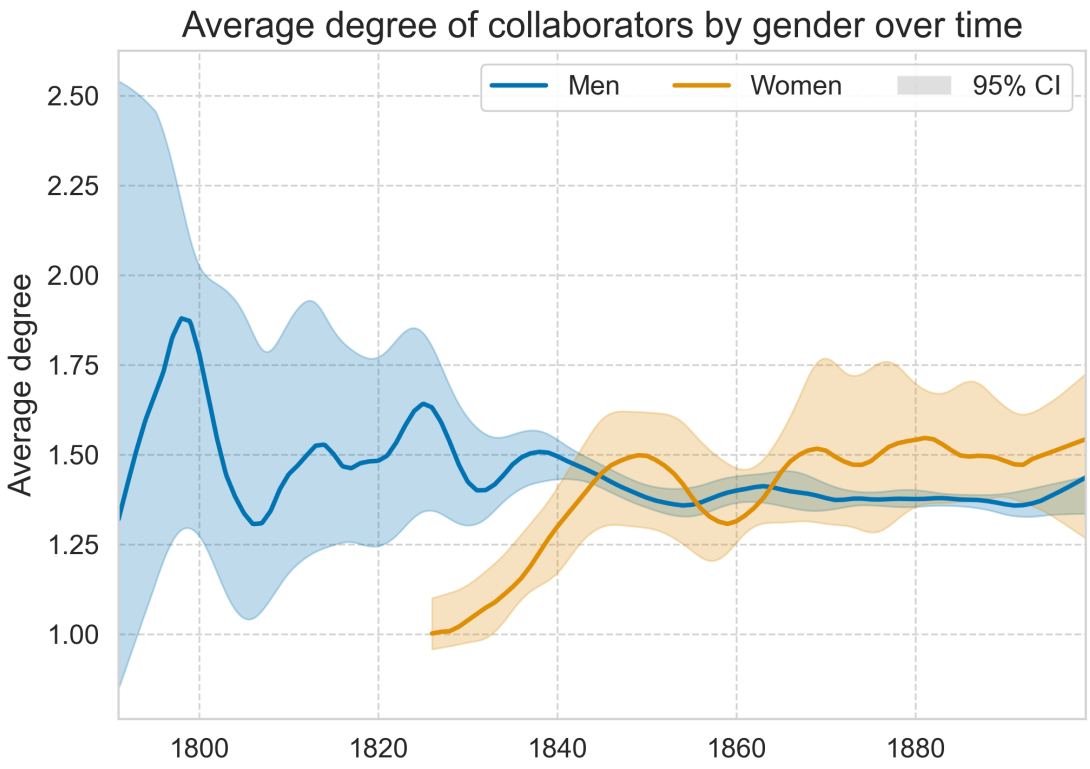


Figure 8: Average number of collaborators per inventor by gender (LOWESS smoothed). Shaded bands denote 95% confidence intervals.

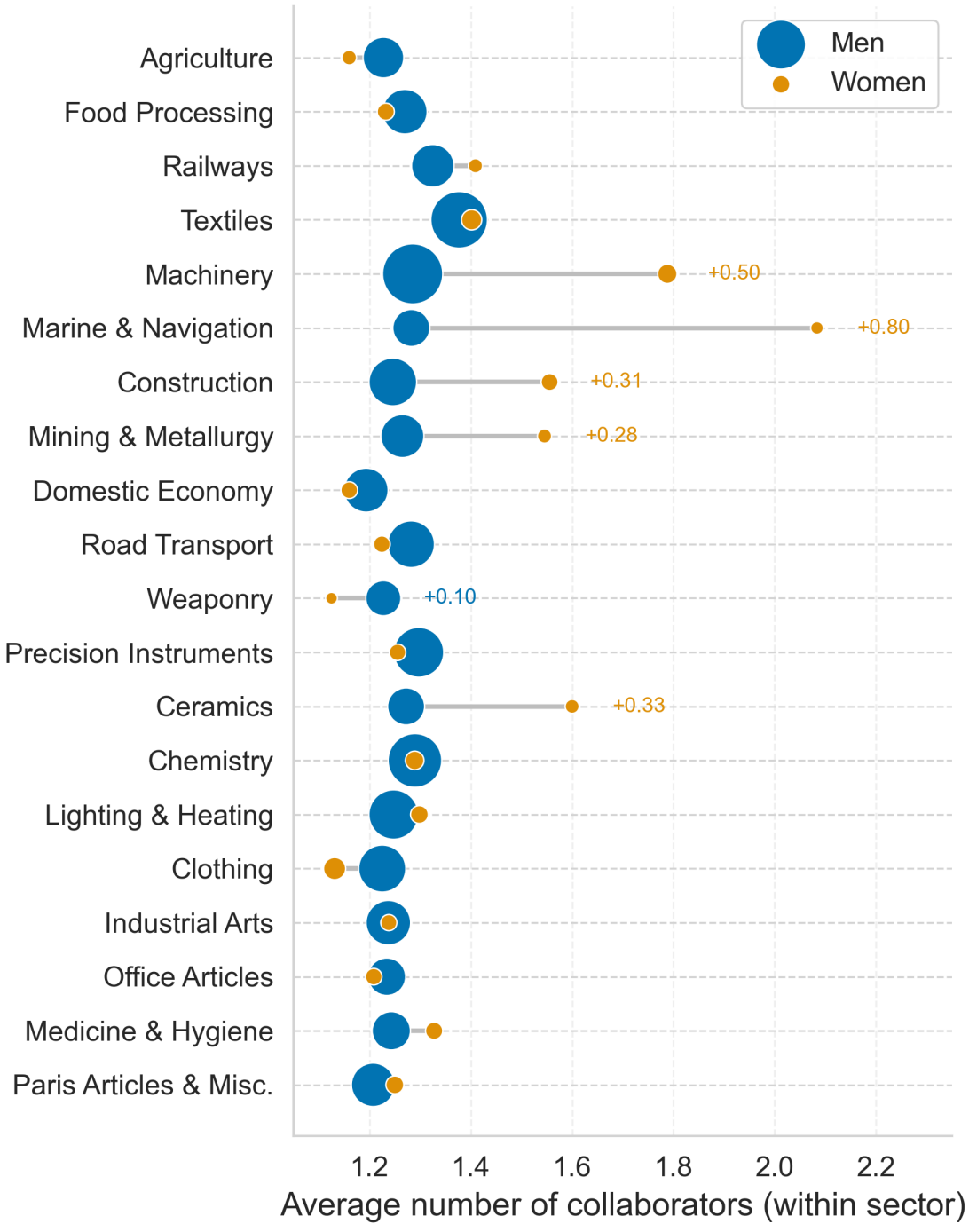


Figure 9: Average number of co-inventors by gender and technology class.

6.3 How do Networks Relate to Diversification?

The event study documented a sharp, persistent increase in inventors' work outside their pre-collaboration home technology class following first collaboration. This section asks which network features are associated with such boundary-crossing more generally. Drawing on the structural holes and cohesion literature reviewed earlier (Burt, 2004; Granovetter, 1973; Reagans and McEvily, 2003), I examine two theoretically motivated mechanisms alongside a benchmark peer-exposure channel.

The literature suggests two complementary pathways through which networks enable boundary-crossing. First, *range*, ties into diverse pools, is associated with access to non-redundant knowledge and alternative problem framings (Burt, 2004; Reagans and McEvily, 2003). Inventors with collaborators from different technology classes can more easily see and pursue opportunities outside their home field. Second, *cohesion*, repeated, trust-laden ties, facilitates the transfer of tacit knowledge needed to actually work in a new domain (Fleming et al., 2007b; Reagans and McEvily, 2003). Crossing a technological boundary once with a partner builds shared understanding; repeating with that same partner reinforces the capacity to operate outside the home field. As a benchmark, I also test whether simply observing neighbors' boundary-crossing activity matters, a weaker form of knowledge transfer than direct collaboration.

To broaden the sample beyond event-study solo-to-collaboration switchers, I redefine the home technology class as each inventor's entry class, the technology class of their first (main) patent, C_i^{entry}, which all inventors possess by construction. The outcome is the inventor–year share of main patents filed outside that entry class:

$$\text{OutsideClassShare}_{it} = \frac{\text{Number of main patents of inventor } i \text{ in year } t \text{ outside } C_i^{\text{entry}}}{\text{Number of main patents of inventor } i \text{ in year } t}.$$

I construct the collaboration network cumulatively by year, defining an edge between inventors i and n in the first year they co-patent, then persisting it forward. All regressors use information through $t-1$ only to avoid simultaneity. I estimate inventor–year regressions with inventor and year fixed effects, standardize all covariates, and double-cluster standard errors by inventor and year. The sample contains 119,412 inventor-years from 41,804 inventors (1791–1900) with positive annual output, at least two career patents, and a defined entry class. Multicollinearity is modest: all variance inflation factors are below 3.6.

I measure an inventor's *range* at $t-1$ as the share of their collaborators whose own entry class differs from C_i^{entry}:

$$\text{CrossClassPartnerShare}_{i,t-1} = \frac{1}{|N_{i,t-1}|} \sum_{n \in N_{i,t-1}} \mathbf{1}(C_n^{\text{entry}} \neq C_i^{\text{entry}}),$$

where $N_{i,t-1}$ is the set of collaborators accumulated through $t-1$. This captures the structural-holes logic: a larger menu of partners from different fields lowers the search cost of stepping outside (Burt, 2004). I control for four network features that might confound this relationship: Degree (number of collaborators), NonRedundancy (one minus the density of ties among an inventor's collaborators, capturing structural holes in a different way), FieldReach (the count of distinct technology classes any collaborator has worked in through $t-1$), and FamilyTieShare (the fraction of kin-based collaborations).

Following Reagans and McEvily (2003) and Fleming et al. (2007b), I distinguish repeated ties by whether the pair has already crossed a technological boundary together. A tie to collaborator n is *repeated* at $t-1$ if the pair has filed at least two joint patents by then. I define:

- $R_{in}^{\text{out}}(t-1)$: indicator that the pair (i,n) is repeated by $t-1$ and has jointly patented outside C_i^{entry} by $t-1$
- $R_{in}^{\text{home}}(t-1)$: indicator that the pair (i,n) is repeated by $t-1$ but has never jointly patented outside C_i^{entry} by $t-1$

I then compute:

$$\text{RepeatedOutsideTieShare}_{i,t-1} = \frac{1}{|N_{i,t-1}|} \sum_{n \in N_{i,t-1}} R_{in}^{\text{out}}(t-1),$$

$$\text{RepeatedHomeTieShare}_{i,t-1} = \frac{1}{|N_{i,t-1}|} \sum_{n \in N_{i,t-1}} R_{in}^{\text{home}}(t-1).$$

The first measure captures cohesion paired with prior boundary-crossing—the partner with whom one has already ventured outside once. The second captures cohesion without such history. If cohesion enables tacit knowledge transfer and trust (Reagans and McEvily, 2003), then repeating with a partner *after* an initial crossing should reinforce diversification, while repeating only within the home field should not.

As a benchmark, I construct $\text{NbrOutputOutside}_{i,t-1}$, the average share of each collaborator's own patents (in year $t-1$) that fall outside C_i^{entry}. This tests whether simply watching collaborators work outside one's home field spills over, independent of direct joint projects. The literature suggests such observational effects should be weaker than direct collaboration (Reagans and McEvily, 2003).

Table 1: Panel Regressions of Network and Diversification

	Dependent variable: Share of Patents Outside Entry Class (t)		
	Cross-Class Ties	Neighbor Exposure	Combined
	(1)	(2)	(3)
Cross-class Partner Share (t-1)	0.121***		0.121***
	(0.005)		(0.005)
x Female	0.092***		0.092***
	(0.023)		(0.023)
Degree (t-1)	-0.058***	-0.025***	-0.058***
	(0.008)	(0.006)	(0.008)
Non-redundancy (t-1)	0.017***	0.017***	0.017***
	(0.004)	(0.003)	(0.004)
Class Reach (t-1)	-0.039***	-0.002	-0.041***
	(0.007)	(0.008)	(0.007)
Family Tie Share (t-1)	-0.027***	-0.003	-0.027***
	(0.006)	(0.006)	(0.006)
Repeated Outside-Tie Share (t-1)	0.060***	0.076***	0.059***
	(0.004)	(0.004)	(0.004)
x Female	0.065**	0.063***	0.062**
	(0.023)	(0.017)	(0.023)
Repeated Home-Only Tie Share (t-1)	-0.007*	-0.009**	-0.007*
	(0.003)	(0.003)	(0.003)
Neighbors' Outside Output (t-1)		0.005*	0.003
		(0.002)	(0.002)
x Female		0.008	0.012
		(0.016)	(0.016)
Inventor FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	119412	119412	119412
N. of groups	41804	41804	41804
Within R^2	0.067	0.033	0.067

Note: *p<0.05; **p<0.01; ***p<0.001

All regressors are standardized (mean 0, SD 1). Dependent variable is the inventor-year share of main patents outside the inventor's entry class. All models include inventor and year fixed effects; SEs are double-clustered by inventor and year. Sample: inventor-years with positive output, ≥ 2 career main patents, 1791–1900. Multicollinearity is modest (all VIFs < 3.6).

Table 1 presents three specifications. Model 1 includes range (CrossClassPartnerShare) and its interaction with gender. Model 2 includes peer exposure (NbrOutputOutside) and its interaction. Model 3 combines both mechanisms. All models include the cohesion measures (RepeatedOutside-

TieShare, RepeatedHomeTieShare) and controls, with the cohesion-gender interaction included throughout.

Range shows the strongest association. A one-standard-deviation increase in $\text{CrossClassPartnerShare}_{i,t-1}$ raises $\text{OutsideClassShare}_{it}$ by 12.1 percentage points (s.e. 0.005). The single largest predictor in the model. This aligns with the structural-holes logic: access to collaborators from different fields directly enables boundary-crossing (Burt, 2004).

Cohesion operates asymmetrically. A one-SD increase in $\text{RepeatedOutsideTieShare}_{i,t-1}$, the share of repeated ties with prior joint boundary-crossing, raises diversification by 6 percentage points (s.e. 0.004). In contrast, $\text{RepeatedHomeTieShare}_{i,t-1}$, repeated ties that have never left the home field, shows a small negative association of -0.007 (s.e. 0.003). Repetition per se does not diversify; repetition after an initial crossing with that same partner does. This fits the cohesion-and-tacit-transfer framework: the trusted relationship formed during an initial joint project outside the home field makes subsequent boundary-crossing with that partner easier (Fleming et al., 2007b; Reagans and McEvily, 2003).

Peer exposure is weak. $\text{NbrOutputOutside}_{i,t-1}$ shows a small, statistically marginal association of approximately 0.003 ($p \approx 0.06$) once range and cohesion are included (Model 3). Observing collaborators' boundary-crossing behavior matters far less than directly partnering with them or repeating with partners who have already helped make a crossing. This result is consistent with the finding that tacit knowledge requires direct collaboration rather than observation (Reagans and McEvily, 2003).

The controls behave as expected. NonRedundancy (structural holes within one's local network) is positively associated with crossing ($\approx +0.017$), consistent with Burt's account of brokerage (Burt, 2004). In contrast, Degree (≈ -0.058), FieldReach (≈ -0.041), and FamilyTieShare (≈ -0.027) negatively predict crossing. This could suggest that inventors with many collaborators, broad cumulative class histories, or kin-based networks tend to remain anchored in their entry field, possibly due to accumulated obligations or specialization.

The effects of range and cohesion are substantially larger for women. The CrossClassPartnerShare $\times$ Female interaction is approximately 0.092 (s.e. 0.023), and the RepeatedOutsideTieShare $\times$ Female interaction is approximately 0.062 (s.e. 0.023). In contrast, the peer-exposure interaction (NbrOutputOutside $\times$ Female) is small and statistically insignificant (≈ 0.012, s.e. 0.016). These patterns align with research showing that when institutional barriers are higher, direct access to diverse, trusted partners becomes especially valuable for crossing boundaries (Ling et al., 2025; Meng, 2016). Women face higher thresholds for legitimacy; the trusted cross-field tie or the repeated partnership with someone who has already helped make a crossing lowers those thresholds more than passive observation of peers' activity.

The network evidence complements the event study. Collaboration reallocates inventive effort across technology according to channels grounded in the structural holes and cohesion literature: access to collaborators from different fields (range) and reinforcement through repeated partnerships with those who have already helped cross a boundary (cohesion). Observing peers' boundary-crossing contributes little once direct partnership is accounted for. These mechanisms are stronger for women, consistent with higher barriers to technological boundary-crossing and greater reliance on trusted, cross-field ties to overcome them.

7 Discussion and Conclusion

When inventors in nineteenth-century France moved from solo work to collaboration, they did not become more productive or create more original work. They did initially produce work that was more influential to later inventors. What they did more consistently however was reallocate effort across technology classes. The share of output filed outside an inventor's home class jumped by roughly eleven percentage points at first collaboration and remained elevated for years. This diversification was partly enabled by network position: access to collaborators from different technology classes opened the door, and repeated partnerships with those who had already helped cross a technology boundary kept it open. Women, who were scarce and relied heavily on family ties for access, extracted substantially larger returns from cross-class partnerships when they obtained them.

The finding of a positive influence but absence of productivity and novelty gains requires explanation. Wuchty et al. (2007) show that teams increasingly dominate the upper tail of the citation distribution across science and patenting by the late twentieth century. But they also show that this change is recent. In the 1950s solo authors in science and engineering and the social sciences were more likely than teams to receive the highest number of citations. Their patent data starts later, in the mid 1970s and is more in line with the dominance of teams. The temporary increase in influence, akin to citations, shortly following the move to collaboration could be explained by teams being better at diffusing their ideas than solo collaborators, or that their work is simply of higher quality depending on the depth of knowledge in certain sectors (Bloom et al., 2020; Jones, 2009).

The work of Uzzi et al. (2013) shows that the most influential scientific papers rely on atypical combinations embedded within otherwise conventional work, and that teams are markedly better than solo authors at achieving this balance. The influence patterns in this study are consistent with that view: collaboration is associated with more influential output (see also robustness in Appendix F). Yet this does not translate into higher novelty. The null results on novelty hold even when distinguishing between system-wide novelty and novelty measured across or within technology classes (see Appendix E.2). If anything, the only significant effect appears in the within technology class, where collaboration is associated with lower novelty. One interpretation is that novelty, as an outcome, reflects a different process than influence. Wu et al. (2019) find that solo authors are substantially more likely to produce highly disruptive work than large teams. Disruptivity and influence therefore need not move together: teams may produce work that diffuses widely and shapes subsequent developments without significantly departing from existing practice. This tension, and the balance between conventionality and atypicality, helps explain why collaboration in nineteenth-century France temporarily increased influence without raising productivity or originality. It may be that inventors' first collaborative patents managed that balance especially well.

The diversification results fit closely with network-based accounts of how ideas move across fields. Burt (2004) argues that actors who bridge structural holes gain access to non-redundant knowledge and alternative frames. In this study, collaborators from other technology classes play that role. The event study shows that first collaboration coincides with a large and persistent expansion of work outside the home class. The network regressions then clarify how this happens. Cross-class partner share is the strongest predictor of boundary-crossing, and repeated ties with partners who have already helped file outside the home class are also strongly associated with continued diversification. Simply having many collaborators, or watching collaborators diversify on their own, matters far less. This pattern echoes Reagans and McEvily (2003) and Fleming et al. (2007b) finding that tacit knowledge and recombination travel best through direct, trusted collaboration rather than through observation alone.

The gender results show how access to such ties was uneven. Women were less numberous than men but were more likely to patent in teams (Merouani and Perrin, 2026). At the same time, they relied heavily on family links to enter inventive work at all. Those kin ties provided access but often kept activity close to existing lines of work. When women did obtain cross-class collaborators or repeated ties that had already produced patents outside their home class, the association with diversification was substantially larger than for men. This is in line with work that stresses higher thresholds of trust and proof for women in networks rich in structural holes (Burt, 1998; Ling et al., 2025; Meng, 2016). In a setting where women faced strong barriers to entry, the rare trusted partner from another field could help realize what they wanted to work on. This could be especially important in the very fields were men dominated the technology class.

These findings come with limitations. Patents record formal credit, not the full social process of invention, so collaboration may mix genuine joint problem solving with financial partnerships, hierarchical relationships, or family arrangements. The network analysis is observational and cannot fully separate selection into certain kinds of ties from their effects. The novelty and influence measures are text based, derived only from patent titles, and rely on the assumption that semantic distance maps to conceptual originality and impact of work. Finally, the focus on inventors with at least two patents omits one-time patentees, who may have different collaboration patterns and returns.

Even with these caveats, the results speak to broader debates about collaboration, innovation and the openness of the French patent system. Methodologically, the text-based measures show that it is possible to study novelty and influence in historical settings without citation data. Substantively,

the study suggests that the gains from collaboration depend on context. In modern science and technology, where knowledge is deep, codified, and widely shared, teams often matter because they allow specialization and support highly novel combinations (Jones, 2009; Uzzi et al., 2013; Wu et al., 2019). In nineteenth-century France, collaboration mainly changed where inventors applied their skills, not how much they produced or how original their ideas were. It worked as a tool for crossing technological boundaries, especially when it linked inventors to partners from other fields and when women used scarce, trusted ties to enter new areas. The dispersed network structure, with few brokers and no dominant hubs, fits this interpretation and connects back to the question of the openness of the patent system in France: rather than a highly centralized, elite-dominated system, the pattern of many small, weakly connected teams is more consistent with a patent regime that, despite its costs, remained relatively open to a broad set of inventors.

References

Bafna, P., Pramod, D., and Vaidya, A. 2016. Document clustering: TF-IDF approach, pp. 61–66, in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*

Barcina-Blanco, M., Lobo, J. L., Garcia-Bringas, P., and Del Ser, J. 2024. Managing the unknown in machine learning: Definitions, related areas, recent advances, and prospects, *Neurocomputing*, vol. 599, 128073

Baudry, J. 2015. L'échec et les brevets d'invention (France, 1791-1844), pp. 47–56, in Coquery, N. and de Oliveira, M. (eds.), *L'échec a-t-il des vertus économiques*, Vincennes, Institut de la gestion publique et du développement économique

Baudry, J. 2019. Examining inventions, shaping property: The savants and the French patent system, *History of Science*, vol. 57, no. 1, 62–80

Blalock, H. M. 1967. *Toward a theory of minority-group relations*, New York, N.Y., Wiley

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. 2020. Are Ideas Getting Harder to Find?, *American Economic Review*, vol. 110, no. 4, 1104–44

Burt, R. S. 1992. *Structural Holes: The Social Structure of Competition*, https://papers.ssrn.com/abstract=1496205 (date last accessed 11 November 2025)

Burt, R. S. 1998. The Gender of Social Capital, *Rationality and Society*, vol. 10, no. 1, 5–46

Burt, R. S. 2004. Structural Holes and Good Ideas., *American Journal of Sociology*, vol. 110, no. 2, 349–99

Chan, T. H., Mihm, J., and Sosa, M. 2021. Revisiting the Role of Collaboration in Creating Breakthrough Inventions, *Manufacturing & Service Operations Management*, vol. 23, no. 5, 1005–24

Clapham, J. H. 1921. *The economic development of France and Germany, 1815-1914*, Cambridge [Eng.] The University press

Ding, W. W., Murray, F., and Stuart, T. E. 2006. Gender differences in patenting in the academic life sciences, *Science (New York, N.Y.)*, vol. 313, no. 5787, 665–67

Emptoz, G. and Marchal, V. 2002. *Aux sources de la propriété industrielle: guide des archives de l'INPI*, INPI

Enamorado, T., Fifield, B., and Imai, K. 2019. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records, *American Political Science Review*, vol. 113, no. 2, 353–71

Fleming, L. 2001. Recombinant Uncertainty in Technological Search, *Management Science*, vol. 47, no. 1, 117–32

Fleming, L., King, C., and Juda, A. I. 2007a. Small Worlds and Regional Innovation, *Organization Science*, vol. 18, no. 6, 938–54

Fleming, L., Mingo, S., and Chen, D. 2007b. Collaborative Brokerage, Generative Creativity, and Creative Success, *Administrative Science Quarterly*, vol. 52, no. 3, 443–75

Galvez-Behar, G. 2019. The patent system during the French industrial revolution: Institutional change and economic effects, *Jahrbuch für Wirtschaftsgeschichte/Economic History Yearbook*, vol. 60, no. 1, 31–56

Garicano, L. 2000. Hierarchies and the Organization of Knowledge in Production, *Journal of Political Economy*, vol. 108, no. 5, 874–904

Granovetter, M. S. 1973. The Strength of Weak Ties, *American Journal of Sociology*, vol. 78, no. 6, 1360–80

Griliches, Z. 1990. Patent Statistics as Economic Indicators: A Survey, *Journal of Economic Literature*, vol. 28, no. 4, 1661–1707

Hall, B., Jaffe, A., and Trajtenberg, M. 2001. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools: National Bureau of Economic Research w8498, w8498 p., date last accessed October 7, 2025, at http://www.nber.org/papers/w8498.pdf

Hall, B. H., Jaffe, A., and Trajtenberg, M. 2005. Market Value and Patent Citations, *The RAND Journal of Economics*, vol. 36, no. 1, 16–38

Hanson, S. 2010. *Lost Talent*, Labor And Social Change, Philadelphia, Temple University Press

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. 1999. Citation Frequency and the Value of Patented Inventions, *Review of Economics and Statistics*, vol. 81, no. 3, 511–15

Hirsch, J.-P. and Minard, P. 1998. 'Laissez-nous faire et protégez-nous beaucoup': Pour une histoire des pratiques institutionnelles dans l'industrie française, XVIII –XIX siècles, pp. 135–58, in Bergeron, L. and Bourdelais, P. (eds.), *La France n'est-elle pas douée pour l'industrie?*, Paris, Belin

Jaffe, A. B. 1986. Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value, *The American Economic Review*, vol. 76, no. 5, 984–1001

Jaffe, A. B. 1989. Real Effects of Academic Research, *The American Economic Review*, vol. 79, no. 5, 957–70

Jones, B. F. 2009. The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?, *Review of Economic Studies*, vol. 76, no. 1, 283–317

Jung, T. and Ejermo, O. 2014. Demographic patterns and trends in patenting: Gender, age, and education of inventors, *Technological Forecasting and Social Change*, vol. 86, 110–24

Kanter, R. M. 2010. *Men and women of the corporation*, New York, NY, Basic Books

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. 2021. Measuring Technological Innovation over the Long Run, *American Economic Review: Insights*, vol. 3, no. 3, 303–20

Khan, B. Z. 2005. *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790-1920*, Cambridge University Press

Khan, B. Z. 2020. *Inventing Ideas: Patents, Prizes, and the Knowledge Economy*, Oxford, New York, Oxford University Press

Khan, B. Z. 2024. 'A new way by her invented': Women inventors and technological innovation in Britain, 1800–1930, *The Economic History Review*, vol. 77, no. 3, 928–52

Khan, B. Z. and Sokoloff, K. L. 2008. A Tale of Two Countries: Innovation and Incentives Among Great Inventors in Britain and the United States, 1750–1930, in Farmer, R. E. A. (ed.), *Macroeconomics in the Small and the Large*, Edward Elgar Publishing

Kreutzer, P. J. and Taalbi, J. 2025. Collaboration for the Bioeconomy

La Mela, M., Frankemölle, J., and Tell, F. 2024. Novelty and impact: Using document similarity to study important inventions in historical Swedish patents, 1890–1929, in Digital Humanities in the Baltic and Nordic Countries 8th Conference –

Landes, D. S. 1949. French Entrepreneurship and Industrial Growth in the Nineteenth Century, *The Journal of Economic History*, vol. 9, no. 1, 45–61

Landes, D. S. 1969. *The unbound Prometheus: Technological change and industrial development in Western Europe from 1750 to the present*, Cambridge, U.K. :: Cambridge University Press Ann Arbor, Michigan: MPublishing, University of Michigan Library

Lanjouw, J. O. and Schankerman, M. 2001. Characteristics of Patent Litigation: A Window on Competition, *The RAND Journal of Economics*, vol. 32, no. 1, 129

Lee, S. and Bozeman, B. 2005. The Impact of Research Collaboration on Scientific Productivity, *Social Studies of Science*, vol. 35, no. 5, 673–702

Li, Y., Li, J., Suhara, Y., Doan, A., and Tan, W.-C. 2023. Effective entity matching with transformers, *The VLDB Journal*, vol. 32, no. 6, 1215–35

Lin, N. 2001. *Social Capital: A Theory of Social Structure and Action*, Cambridge University Press

Ling, J., Brands, R., Brass, D. J., Liu, D., Borgatti, S. P., and Mehra, A. 2025. Gender, Structural Holes, and Citations: The Effects of Women's Increasing Proportional Representation in a Field, *Group & Organization Management*, 10596011251326317

MacLeod, C. and Nuvolari, A. 2016. Inventive Activities, Patents and Early Industrialisation: A Synthesis of Research Issues, *Rivista di storia economica*, no. 1, 77–108

Marchal, V. 2009. Brevets, marques, dessins et modèles. Évolution des protections de propriété industrielle au XIXe siècle en France, *Documents pour l'histoire des techniques*, no. 17, 106–16

Meng, Y. 2016. Collaboration patterns and patenting: Exploring gender distinctions, *Research Policy*, vol. 45, no. 1, 56–67

Merouani, Y. and Perrin, F. 2026. Women Inventors: On the Origins of the Gender Patenting Gap, *Journal of Economic History*, Advance Access published 2026

Nuvolari, A. and Tartari, V. 2011. Bennet Woodcroft and the value of English patents, 1617–1841, *Explorations in Economic History*, vol. 48, no. 1, 97–115

Nuvolari, A., Tortorici, G., and Vasta, M. 2023. British-French Technology Transfer from the Revolution to Louis Philippe (1791–1844): Evidence from Patent Data, *The Journal of Economic History*, vol. 83, no. 3, 833–73

O'Brien, P. and Keyder, Ç. 1978. *Economic growth in Britain and France 1780 - 1914: Two paths to the twentieth century*, London, Allen & Unwin

Pakes, A. 1986. Patents as Options: Some Estimates of the Value of Holding European Patent Stocks, *Econometrica*, vol. 54, no. 4, 755

Perry-Smith, J. E. 2006. Social Yet Creative: The Role Of Social Relationships In Facilitating Individual Creativity, *Academy of Management Journal*, vol. 49, no. 1, 85–101

Reagans, R. and McEvily, B. 2003. Network Structure and Knowledge Transfer: The Effects of Cohesion and Range, *Administrative Science Quarterly*, vol. 48, no. 2, 240–67

Roehl, R. 1976. French industrialization: A reconsideration, *Explorations in Economic History*, vol.

13, no. 3, 233–81

Rossiter, M. W. 1995. *Women scientists in America : Before affirmative action, 1940-1972*, Baltimore : Johns Hopkins University Press

Schankerman, M. and Pakes, A. 1986. Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period, *The Economic Journal*, vol. 96, no. 384, 1052

Serrano, C. J. 2010. The dynamics of the transfer and renewal of patents, *The RAND Journal of Economics*, vol. 41, no. 4, 686–708

Singh, J. and Fleming, L. 2009. Lone Inventors as Source of Breakthroughs: Myth or Reality?, *Management Science*, Advance Access published 10 June 2009: doi:10.1287/mnsc.1090.1072

Sun, L. and Abraham, S. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics*, vol. 225, no. 2, 175–99

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. 2013. Atypical Combinations and Scientific Impact, *Science*, vol. 342, no. 6157, 468–72

Walton, W. 1992. *France at the Crystal Palace: Bourgeois Taste and Artisan Manufacture in the Nineteenth Century*, Berkeley, University of California Press

Weitzman, M. L. 1998. Recombinant Growth, *The Quarterly Journal of Economics*, vol. 113, no. 2, 331–60

Winkler, W. E. 1988. Using the EM algorithm for weight computation in the fellegi-sunter model of record linkage, pp. 667–71, in *Proceedings of the section on survey research methods*, American Statistical Association

Wuchty, S., Jones, B. F., and Uzzi, B. 2007. The Increasing Dominance of Teams in Production of Knowledge, *Science*, vol. 316, no. 5827, 1036–39

Wu, L., Wang, D., and Evans, J. A. 2019. Large teams develop and small teams disrupt science and technology, *Nature*, vol. 566, no. 7744, 378–82

Yoder, J. D. 1991. Rethinking Tokenism: Looking Beyond Numbers, *Gender & Society*, vol. 5, no. 2, 178–92

Zeakis, A., Papadakis, G., Skoutas, D., and Koubarakis, M. 2025. An in-depth analysis of pre-trained embeddings for entity resolution, *The VLDB Journal*, vol. 34, no. 1, 5

A Record linking

A.1 Extracting Keywords from Patent Titles

To distinguish if two patents fundamentally work with the same product or process despite having different technology classification, I extract keywords from the titles of patents in the French Patent Database, $T = \{t_1, t_2, \ldots, t_N\}$. Each title is first reduced to its essential terms by lemmatizing words and removing stopwords, punctuation, and named entities, producing a refined version t'_i. I then apply a Term Frequency–Inverse Document Frequency (TF–IDF) weighting, which highlights the words that appear often in a given title but rarely across the rest of the corpus [Bafna et al. (2016)]. Formally,

$$\text{TF-IDF}(w, t'_i, T) \;=\; TF(w, t'_i) \times IDF(w, T),$$

where

$$TF(w, t'_i) \;=\; \frac{f_{w,t'_i}}{\sum_{w' \in t'_i} f_{w',t'_i}} \qquad \text{and} \qquad IDF(w, T) \;=\; \log\!\left(\frac{|T|}{|\{\, t \in T : w \in t \,\}|}\right).$$

From each title, I then select the highest-weighted words, $K_i = \{w_1, w_2, \ldots, w_k\}$. These keywords summarize the central focus of the invention and provide a way to compare patents on the basis of the vocabulary they share. In my linking strategy, overlapping keywords serve as evidence that two observations may belong to the same underlying entity.

A.2 Blocking rules: Generating candidate pairs

For each type of entity, whether male inventors, women inventors, family firms, or non-family firms, I construct blocking rules to generate candidate record pairs. These rules restrict attention to observations that share plausibly identifying features, which makes the task computationally manageable. I have designed the rules such as they should minimize the risk of missed matches, especially so for the female and male inventors. Their design reflects the attributes most useful for distinguishing entities, such as names, occupations, and technological domains.

Field	Description
patent_code	**Identifier.** The code of the patent family associated with the inventor record (available as `patent_code_l` / `patent_code_r`).
patent_id	**Identifier.** The identifier of the patent application associated with the inventor record (available as `patent_id_l` / `patent_id_r`).
dep_num	**Numeric field.** The applicant order number associated with the observation (available as `dep_num_l` / `dep_num_r`).
app_year	**Numeric field.** The application year of the patent (available as `app_year_l` / `app_year_r`).
surname	**Text field.** The surname of the inventor. Functions as firm-name for firms (available as `surname_l` / `surname_r`).
maiden_name	**Text field.** The maiden name of the inventor (available as `maiden_name_l` / `maiden_name_r`).
first_name	**Text field.** The first name(s) of the inventor (available as `first_name_l` / `first_name_r`).
gender	**Categorical variable.** The gender of the inventor. Values: `"F"` (female), `"M"` (male) (available as `gender_l` / `gender_r`).
mar_status	**Categorical variable.** The marital status of women inventors. Values: `married`, `widow`, `unmarried` (available as `mar_status_l` / `mar_status_r`).
mother	**Binary flag.** `"1"` if an individual shares names with their progeny. Used to separate parent–child entities (available as `mother_l` / `mother_r`).
father	**Binary flag.** `"1"` if an individual shares names with their progeny. Used to separate parent–child entities (available as `father_l` / `father_r`).

Field	Description
married	**Binary flag.** `"0"` if a woman has marital status `unmarried`, `"1"` otherwise (available as `married_l` / `married_r`).
hisco	**Categorical code.** The HISCO code for the inventor's occupation (available as `hisco_l` / `hisco_r`).
keywords	**Array of text.** A list of technical keywords extracted from the patent description (available as `keywords_l` / `keywords_r`).
tech_class	**Categorical variable.** The major technology classification of the patent, 20 categories (available as `tech_class_l` / `tech_class_r`).
tech_minor_class	**Categorical variable.** The minor technology classification of the patent, 97 categories (available as `tech_minor_class_l` / `tech_minor_class_r`).
place_id_list	**Array of identifiers.** A list of place IDs (city-level) associated with the patent (available as `place_id_list_l` / `place_id_list_r`).

A.2.1 Firm inventor blocking rules

This section describes the blocking rules used for firm inventors, whether family-owned or otherwise. Each rule requires agreement on a central identifier (such as surname) and adds restrictions on time, occupation, technology classification, or technical vocabulary. I exclude cases where two records refer to the same patent application. Table Table 3 summarizes the blocking conditions. Note that the formal conditions mirror the syntax used in the entity resolution implementation.

Table 3: Firm inventor blocking rules

Rule	Description	Formal Condition
1	**Same name and occupation.** Two firms are compared if they share a surname and HISCO occupation code, apply for patents within 50 years of each other, and are associated with distinct patent families and applications.	`ABS(l.app_year - r.app_year) < 50 and l.surname = r.surname and l.hisco = r.hisco and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.gender = r.gender`
2	**Same name and shared technical keyword.** Firms are compared if at least one technical keyword from the patent description overlaps, provided applications are within 50 years and do not belong to the same patent application or family.	`ABS(l.app_year - r.app_year) < 50 and l.surname = r.surname and (list_unique(l.keywords) + list_unique(r.keywords) - list_unique(list_concat(l.keywords, r.keywords)) >= 1) and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.gender = r.gender`
3	**Same name and technology class.** Comparisons are allowed when firms share a surname and technology class, with applications within 50 years and distinct patents.	`ABS(l.app_year - r.app_year) < 50 and l.surname = r.surname and l.tech_class = r.tech_class and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.gender = r.gender`
4	**Within patent-family differentiation.** If firms share a surname and depositor number within the same patent family, but correspond to different applications, they are compared.	`l.surname = r.surname and l.dep_num = r.dep_num and l.patent_code = r.patent_code and l.patent_id != r.patent_id and l.gender = r.gender`

Rule	Description	Formal Condition
5	**Female-linked firms only.** For firms associated with female inventors, I allow comparisons across distinct patents whenever surnames coincide.	`l.surname = r.surname and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.gender = r.gender and l.gender = 'F'`

A.2.2 Female inventor blocking rules

For female inventors, I use six rules. The rules rely on surnames, first names, maiden names, marital status, and the mother indicator to link records across life-cycle name changes. Table Table 4 summarizes the conditions.

Table 4: Female inventor blocking rules

Rule	Description	Formal Condition
1	**Same surname, marital status, and applicant order.** Compare records when surname, marital status, and `dep_num` match; patent applications differ; and the mother indicator matches.	`l.surname = r.surname and l.mar_status = r.mar_status and l.dep_num = r.dep_num and l.patent_id != r.patent_id and l.mother = r.mother`
2	**Same surname and first name, same married flag.** Compare when surname and first name match and the binary `married` flag is equal; applications differ; mother indicator matches.	`l.surname = r.surname and l.first_name = r.first_name and l.married = r.married and l.patent_id != r.patent_id and l.mother = r.mother`
3	**Same surname and maiden name, same married flag.** Compare when surname and maiden name match and the `married` flag is equal; applications differ; mother indicator matches.	`l.surname = r.surname and l.maiden_name = r.maiden_name and l.married = r.married and l.patent_id != r.patent_id and l.mother = r.mother`
4	**Cross-surname link via maiden name (status changes).** Compare when surnames differ but one record's maiden name equals the other's surname, the `married` flags differ (capturing unmarried vs. married/widow), applications differ, and the mother indicator matches.	`l.surname != r.surname and l.married != r.married and (l.maiden_name = r.surname OR r.maiden_name = l.surname) and l.patent_id != r.patent_id and l.mother = r.mother`
5	**Same surname and marital status.** Compare when surname and marital status match; applications differ; mother indicator matches.	`l.surname = r.surname and l.mar_status = r.mar_status and l.patent_id != r.patent_id and l.mother = r.mother`
6	**Same surname only (residual).** Compare when surnames match; applications differ; mother indicator matches.	`l.surname = r.surname and l.patent_id != r.patent_id and l.mother = r.mother`

A.2.3 Male inventor blocking rules

For male inventors, I use six rules. The rules rely on surnames, first names, occupations, technology classifications, keywords from patent descriptions, applicant order, and place identifiers. To improve match quality, addresses corresponding to Paris (`place_id_list = "641322"`) are excluded, since they occur very frequently. As in other cases, identical patent applications are not compared. Table Table 5 lists the conditions.

Table 5: Male inventor blocking rules

Rule	Description	Formal Condition
1	**Same surname and occupation.** Compare records if surname and HISCO occupation match, applications are within 30 years, patent families differ, applications differ, and the father indicator matches.	`ABS(l.app_year-r.app_year)<30 and l.surname = r.surname and l.hisco = r.hisco and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.father = r.father`
2	**Same surname and shared technical keyword.** Compare when at least one keyword from the patent description overlaps, surname matches, applications are within 30 years, patents differ, and the father indicator matches.	`ABS(l.app_year-r.app_year)<30 and l.surname = r.surname and (list_unique(l.keywords) + list_unique(r.keywords) - list_unique(list_concat(l.keywords, r.keywords)) >= 1) and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.father = r.father`
3	**Same surname and first name.** Compare when surname and first name match, applications are within 30 years, patents differ, and the father indicator matches.	`ABS(l.app_year-r.app_year)<30 and l.surname = r.surname and l.first_name = r.first_name and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.father = r.father`
4	**Same surname and minor technology class.** Compare when surname and minor technology class match, applications are within 30 years, patents differ, and the father indicator matches.	`ABS(l.app_year-r.app_year)<30 and l.surname = r.surname and l.tech_minor_class = r.tech_minor_class and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.father = r.father`
5	**Within patent-family differentiation.** Compare when surname and applicant order (`dep_num`) match within the same patent family, applications are within 30 years, the application IDs differ, and the father indicator matches.	`ABS(l.app_year-r.app_year)<30 and l.surname = r.surname and l.dep_num = r.dep_num and l.patent_code = r.patent_code and l.patent_id != r.patent_id and l.father = r.father`

6	**Same surname and overlapping place IDs (excluding Paris).** Compare when surname matches, applications are within 30 years, at least one place identifier overlaps after excluding Paris (641322), records are not within same patent application nor same patent-family, and the father indicator matches.	`ABS(l.app_year - r.app_year)<30 and l.surname = r.surname and (array_length(array_intersect(filter(l.place_id_list, x -> x <> '641322'), filter(r.place_id_list, x -> x <> '641322'))) >= 1) and l.patent_code != r.patent_code and l.patent_id != r.patent_id and l.father = r.father`

A.3 Data dictionary for male inventor record linking

A.3.1 I. Identifier Columns

These columns provide unique identifiers for each record and the associated patents. The `_l` and `_r` suffixes denote the "left" and "right" sides of the pair, respectively.

Column Name	Description
`unique_id_l` / `unique_id_r`	A unique identifier for each individual observation (inventor record).
`patent_id_l` / `patent_id_r`	The identifier for the patent-application associated with the inventor record.
`patent_code_l` / `patent_code_r`	The code of the patent-family associated with the inventor record.
`patent_number_l` / `patent_number_r`	The numeric part of the patent code. Often sequential in nature and thus provides some additional information - such as the timing of the patent application.

A.3.2 II. Similarity & Match Score Columns

These columns contain various similarity scores and match probabilities, which are the core of the entity resolution model.

Column Name	Description
`sim_gtr`	A similarity score calculated using the `sentence-transformers/gtr-t5-large` model. The similarity is computed as `1 / (1 + Euclidean distance)` between the text embeddings of the two records.
`sim_mini`	A similarity score calculated using the `sentence-transformers/all-MiniLM-L12-v2` model. The similarity is computed as `1 / (1 + Euclidean distance)` between the text embeddings of the two records.
`s_lm`	The maximum value between `sim_gtr` and `sim_mini`. This is a pragmatic combination of the two SBERT models.

Column Name	Description
match_weight	The final match weight calculated by the Splink model. This is the logarithm of the Bayes factor, which represents the strength of evidence for a match. This value is somewhat useful as it seems mostly useful for the easier matches. Probably because the deterministic rules I set were not excellent, I also set it with a recall of "0.09". The recall here is described as "A guess at the recall the deterministic matching rules will attain. i.e. what proportion of true matches will be recovered by these deterministic rules"
match_probability	The probability that the two records are a match, as predicted by the Splink model. This is derived from the match_weight.
p_raw	Uncalibrated match probability from the RoBERTa classifier (sigmoid of the model logits), in [0, 1]. Good for ranking pairs but not a well-calibrated probability; threshold choices may drift across datasets/training runs.
p_iso	Isotonic-calibrated probability derived from p_raw, fitted on the dev set using isotonic regression. Monotonic and non-parametric, typically yields better probability calibration (especially near 0/1). This will be usef for downstream rules and clustering.

A.3.3 III. Columns from the original dataset

These columns are derived from the original dataset and feature engineered by me directly from the original dataset. The idea here is to provide the cleaned original data such as surnames and first names, but also to provide some features that directly help to match/unmatch pairs. For instance the jeune_aine_cross and other family cross columns when "1" means the row should be a non-match. For manual work, the surname and first name, together with app_year and key words in the patent title are very powerful. Unfortunately, we have large missing values in the first name. The profession has a tendency to change over time, and so it is not always very helpful.

A.3.3.1 A. Name-based Features

Column Name	Description
surname_l / surname_r	The surname of the inventor. No missing values.
first_name_l / first_name_r	The first name(s) of the inventor. 33.5% of the observations have missing values.
first_name_array_l / first_name_array_r	The first name(s) split into a list of strings.
jeune_l / jeune_r	A binary flag (1/0) indicating if the first name contains "jeune" (younger).
aine_l / aine_r	A binary flag (1/0) indicating if the first name contains "aine" (elder).
junior_l / junior_r	A binary flag (1/0) indicating if the first name contains "junior".
senior_l / senior_r	A binary flag (1/0) indicating if the first name contains "senior".
pere_l / pere_r	A binary flag (1/0) indicating if the first name contains "pere" (father).
fils_l / fils_r	A binary flag (1/0) indicating if the first name contains "fils" (son).

Column Name	Description
neveu_l / neveu_r	A binary flag (1/0) indicating if the first name contains “neveu” (nephew).
oncle_l / oncle_r	A binary flag (1/0) indicating if the first name contains “oncle” (uncle).
jeune_aine_cross	A binary flag indicating if one record has “jeune” and the other has “aine”, these should never be a match.
junior_senior_cross	A binary flag indicating if one record has “junior” and the other has “senior”, these should never be a match.
fils_pere_cross	A binary flag indicating if one record has “fils” and the other has “pere”, these should never be a match.
neveu_oncle_cross	A binary flag indicating if one record has “neveu” and the other has “oncle”, these should never be a match.
family_marker	A binary flag indicating if any of the family-related keywords are present in the first names.
first_name_edit_dis	The Damerau-Levenshtein edit distance between the first names.
first_name_jw_sim	The Jaro-Winkler similarity between the first names.
first_name_array_num_l / first_name_array_num_r	The number of tokens in the first_name_array.
first_name_common	A categorical variable indicating the quality of the first name match (‘Perfect Match’, ‘Good Match’, ‘Partial Match’, ‘No Match’).
first_name_common_score	A numeric score corresponding to first_name_common (1.0 for perfect, 0.75 for good, 0.5 for partial, 0.0 for no match).
same_first_name	A binary flag (1/0) indicating if the first names are an exact match.
first_name_embedding_similarity	The cosine similarity between the ‘all-MiniLM-L6-v2’ embeddings of the first names.
contains_initials	A binary flag (1/0) indicating if either first name contains a “.” which is indicative of initials.

A.3.3.2 B. Patent and Invention-based Features

Column Name	Description
dep_num_l / dep_num_r	The applicant order number associated with the observation.
addition_l / addition_r	Binary for main patent (0) or addition (1).
addition_combination	A categorical variable indicating the if the two observations are both from main patents ‘Both main’ or both from patent additions ‘Both addition’ or that they are from different patent types ‘Cross’.
same_patent_family	A binary flag (1/0) indicating if patent_code_l == patent_code_r, which means that the two observations are from the same patent family.

Column Name	Description
profession_l / profession_r	The occupation of the inventor. 54.7% of the observations have missing values.
hisco_l / hisco_r	The HISCO code for the inventor's occupation. 54.7% of the observations have missing values.
unique_words_l / unique_words_r	A list of unique words from the patent description. 0.26% of the observations have missing values.
industry_minor_class_l / industry_minor_class_r	The minor industry classification of the patent, 97 categories. 0.01% of the observations have missing values.
industry_class_l / industry_class_r	The major industry classification of the patent, 20 categories. 0.01% of the observations have missing values.
app_year_l / app_year_r	The application year of the patent. No missing values.
app_year_diff / delta_year	The absolute difference between the application years.
patent_number_distance	The absolute difference between the patent numbers.
industry_minor_class_same	A binary flag (1/0) indicating if the minor industry classes are the same.
industry_major_class_same	A binary flag (1/0) indicating if the major industry classes are the same.
co_inventors_l / co_inventors_r	A list of co-inventors on the patent. 75% of the observations have missing values, because most patents are single-inventor.
titre_patent_addition_l / titre_patent_addition_r	The patent title (short description of the invention). No missing values.
co_inventor_similarity	The Jaccard similarity between the sets of co-inventors.
co_inventor_embedding_similarity	The cosine similarity between the 'all-MiniLM-L6-v2' embeddings of the co-inventor lists.
profession_embedding_similarity	The cosine similarity between the 'all-MiniLM-L6-v2' embeddings of the inventor's professions.
titre_patent_addition_embedding_similarity	The cosine similarity between the 'all-MiniLM-L6-v2' embeddings of the patent titles.
unique_word_interaction	The number of intersecting words in unique_words_l and unique_words_r.
year_bin	A categorical bin for delta_year ('0-2', '3-5', '6-10', '11-20', '21+').

A.3.3.3 C. Location-based Features

Column Name	Description
place_id_list_l / place_id_list_r	A list of place IDs (at the city-level) associated with the Patent.
outside_paris_unique_words_l / outside_paris_unique_words_r	Unique words from the address, excluding "Paris" and "France".
nonparis_overlap	A boolean flag indicating if there is at least one overlapping non-Paris place ID between the two records.

A.3.3.4 D. Patent agent Features

Column Name	Description
`mandataire_exists_l` / `mandataire_exists_r`	A binary flag indicating if a "mandataire" (patent agent) exists for the patent.
`mandataire_place_interaction`	Captures the interaction between location evidence and the presence of a patent agent, as an agent's address can make location matching less reliable. It is calculated as `abs(bf_tf_adj_place_id_list - gamma_mandataire_exists)`.

A.3.4 IV. Splink Model generated features (`gamma_`, `tf_`, `bf_`)

These features are generated by the Splink model and provide detailed information about how the `match_weight` was calculated.

- `gamma_` columns represent the comparison vector for a given feature (e.g., `gamma_surname`). The value corresponds to the level of similarity between the left and right records.
- `tf_` columns contain the term frequency of a given feature in the dataset.
- `bf_` columns represent the Bayes factor for a given comparison level.
- `bf_tf_adj_` columns represent the term-frequency adjusted Bayes factor.

Column Name	Description
`gamma_surname`	The comparison vector for the `surname` field. The levels correspond to: 4 = exact match, 3 = Damerau-Levenshtein <= 1, 2 = Jaro-Winkler >= 0.9, 1 = Jaro-Winkler >= 0.8, 0 = else.
`tf_surname_l` / `tf_surname_r`	The term frequency of the surname on the left/right side of the pair.
`bf_surname`	The Bayes factor for the `surname` comparison level.
`bf_tf_adj_surname`	The term-frequency adjusted Bayes factor for the `surname` comparison.
`gamma_custom_first_name_first_name_array`	The comparison vector for the `first_name_array` field. The levels correspond to: 3 = exact match, 2 = array intersect size >= 2, 1 = array intersect size >= 1, 0 = else.
`tf_first_name_array_l` / `tf_first_name_array_r`	Term frequency of the `first_name_array`.
`bf_custom_first_name_first_name_array`	The Bayes factor for the `first_name_array` comparison.
`bf_tf_adj_custom_first_name_first_name_array`	The term-frequency adjusted Bayes factor for the `first_name_array` comparison.
`gamma_hisco`	The comparison vector for the `hisco` (historical occupation) codes. The levels correspond to an exact match on the full code, and then on the first 4, 3, 2, and 1 digits.
`tf_hisco_l` / `tf_hisco_r`	The term frequency of the `hisco` code.
`bf_hisco`	The Bayes factor for the `hisco` comparison.
`bf_tf_adj_hisco`	The term-frequency adjusted Bayes factor for the `hisco` comparison.
`gamma_custom_unique_words_industry_minor_class_industry_class`	A custom comparison vector for `unique_words`, `industry_minor_class`, and `industry_class`. The levels are hierarchical, starting with `unique_words` and falling back to the industry classes.

Column Name	Description
tf_unique_words_l / tf_unique_words_r	The term frequency of the unique_words.
tf_industry_minor_class_l / tf_industry_minor_class_r	The term frequency of the industry_minor_class.
tf_industry_class_l / tf_industry_class_r	The term frequency of the industry_class.
bf_custom_unique_words_industry_minor_class_industry_class	The Bayes factor for the combined unique_words and industry class comparison.
bf_tf_adj_custom_unique_words_industry_minor_class_industry_class	The term-frequency adjusted Bayes factor for the combined unique_words and industry class comparison.
gamma_app_year	The comparison vector for the app_year. The levels correspond to the number of years between the two patent applications.
bf_app_year	The Bayes factor for the app_year comparison.
gamma_place_id_list	The comparison vector for the place_id_list. The levels correspond to an exact match or an intersection of the lists.
tf_place_id_list_l / tf_place_id_list_r	The term frequency of the place_id_list.
bf_place_id_list	The Bayes factor for the place_id_list comparison.
bf_tf_adj_place_id_list	The term-frequency adjusted Bayes factor for the place_id_list comparison.
gamma_mandataire_exists	The comparison vector for the mandataire_exists flag.
tf_mandataire_exists_l / tf_mandataire_exists_r	The term frequency of the mandataire_exists flag.
bf_mandataire_exists	The Bayes factor for the mandataire_exists comparison.
bf_tf_adj_mandataire_exists	The term-frequency adjusted Bayes factor for the mandataire_exists comparison.

A.3.5 IV. Feature Engineered Columns

These columns were created to provide additional signals for the entity resolution.

A.3.5.1 A. Clustering and Blocking Features

Column Name	Description
id_l_frequency / id_r_frequency	The frequency of the unique_id across both the left and right sides of all pairs.
max_cluster_id	The ID of the connected component (cluster) that the record pair belongs to if we chained all the unique_id's that are being compared.
max_cluster_active_years	The difference between the maximum and minimum app_year within the cluster.
max_cluster_nodes	The total number of unique inventor records in the cluster.
likely_single	A boolean flag indicating if the surname block is likely to contain only a single individual, based on fuzzy name clustering.
num_clusters	The number of distinct first name clusters found within the surname block.
active_duration	The difference between the maximum and minimum app_year within the surname block.

Column Name	Description
surname_block_size	The number of unique inventor records within the surname block.
block_band	A categorical bin for surname_block_size ('small', 'medium', 'large').

A.3.5.2 B. Sampling and Bucketing Features

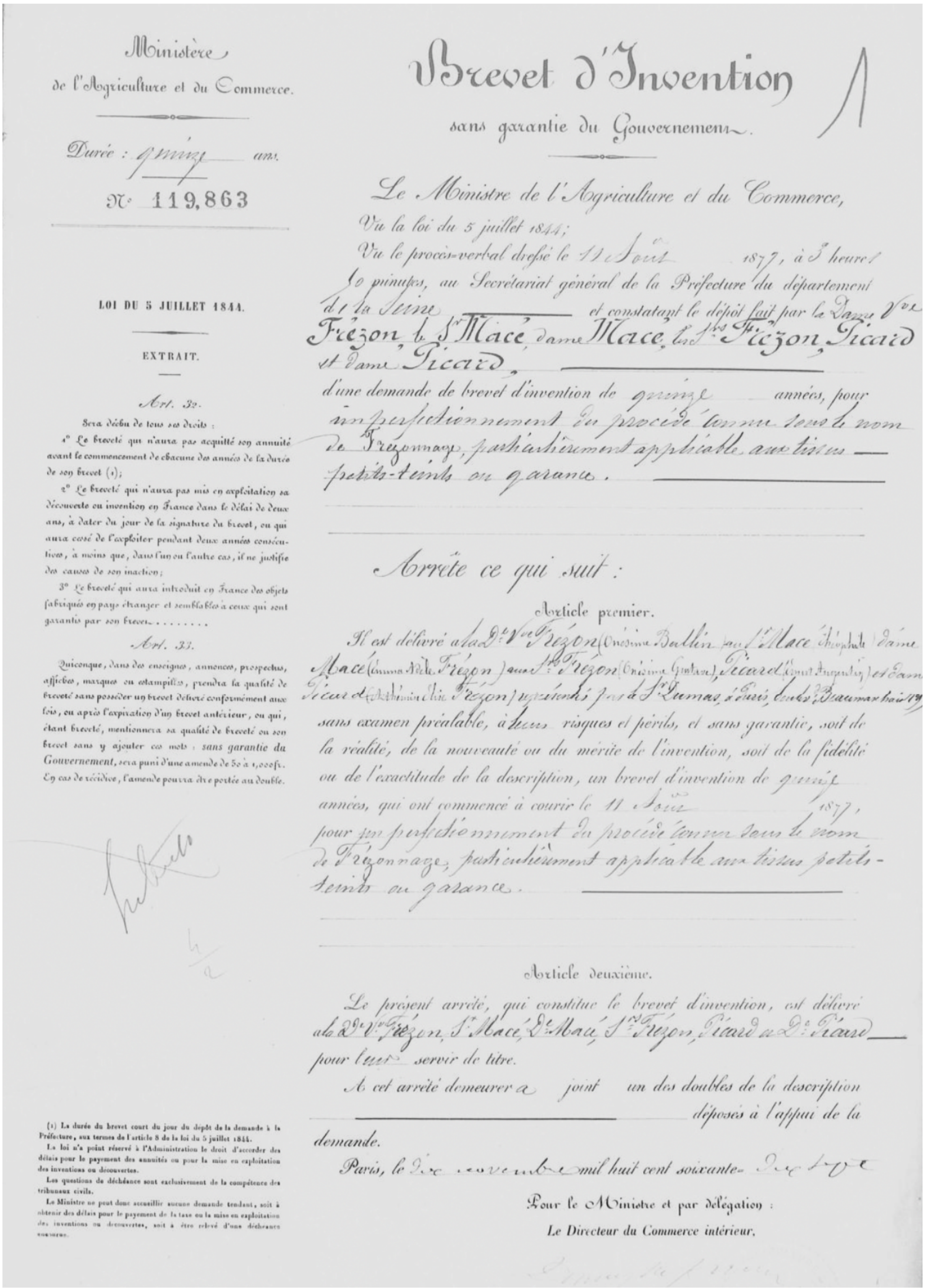
Column Name	Description
fname_cat	A categorical variable for first name availability ('both', 'one', 'none').
minor_match	A binary flag indicating if the minor industry classes are the same.
bucket	Assigns each pair to a sampling stratum based on the agreement or disagreement between the SBERT (s_lm) and Splink (match_probability) scores. The buckets are: **anchor_high**: High agreement (likely match). (s_lm 0.95 or match_probability 0.999). **anchor_low**: High agreement (likely non-match). (s_lm 0.58 and match_probability 0.05). **disagree_sbert_high**: Disagreement. SBERT score is high (s_lm 0.90) while Splink probability is low (match_probability 0.20). **disagree_splink_high**: Disagreement. Splink probability is high (match_probability 0.98) while SBERT score is low (s_lm 0.60). **uncertainty**: SBERT score is in an ambiguous range (0.62 s_lm 0.75).

B Assigning gender to inventors

Work in progress

C Extracting first names using Machine Learning

Work in progress



Ministère
de l'Agriculture et du Commerce.

Durée : quinze ans.

N° 119,863

LOI DU 5 JUILLET 1844.

EXTRAIT.

Art. 32.

Sera déchu de tous ses droits :

1° Le breveté qui n'aura pas acquitté son annuité avant le commencement de chacune des années de la durée de son brevet (1);

2° Le breveté qui n'aura pas mis en exploitation sa découverte ou invention en France dans le délai de deux ans, à dater du jour de la signature du brevet, ou qui aura cessé de l'exploiter pendant deux années consécutives, à moins que, dans l'un ou l'autre cas, il ne justifie des causes de son inaction;

3° Le breveté qui aura introduit en France des objets fabriqués en pays étranger et semblables à ceux qui sont garantis par son brevet........

Art. 33.

Quiconque, dans des enseignes, annonces, prospectus, affiches, marques ou estampilles, prendra la qualité de breveté sans posséder un brevet délivré conformément aux lois, ou après l'expiration d'un brevet antérieur, ou qui, étant breveté, mentionnera sa qualité de breveté ou son brevet sans y ajouter ces mots : sans garantie du Gouvernement, sera puni d'une amende de 50 à 1,000 fr. En cas de récidive, l'amende pourra être portée au double.

(1) La durée du brevet court du jour du dépôt de la demande à la Préfecture, aux termes de l'article 8 de la loi du 5 juillet 1844.
La loi n'a point réservé à l'Administration le droit d'accorder des délais pour le payement des annuités ou pour la mise en exploitation des inventions ou découvertes.
Les questions de déchéance sont exclusivement de la compétence des tribunaux civils.
Le Ministre ne peut donc accueillir aucune demande tendant, soit à obtenir des délais pour le payement de la taxe ou la mise en exploitation des inventions ou découvertes, soit à être relevé d'une déchéance encourue.

Brevet d'Invention

sans garantie du Gouvernement.

Le Ministre de l'Agriculture et du Commerce,

Vu la loi du 5 juillet 1844;

Vu le procès-verbal dressé le 11 Août 1877, à 3 heures 50 minutes, au Secrétariat général de la Préfecture du département de la Seine et constatant le dépôt fait par la Dame Vve Frézon, le Sr Macé, dame Macé, le Sr Frézon, Picard et dame Picard, d'une demande de brevet d'invention de quinze années, pour un perfectionnement du procédé connu sous le nom de Frézonnage, particulièrement applicable aux tissus petits-teints ou garance.

Arrête ce qui suit :

Article premier.

Il est délivré à la Dme Vve Frézon [illegible], au Sr Macé [illegible], dame Macé [illegible], au Sr Frézon [illegible], Picard [illegible] et dame Picard [illegible], représentés par le Sr [illegible], sans examen préalable, à leurs risques et périls, et sans garantie, soit de la réalité, de la nouveauté ou du mérite de l'invention, soit de la fidélité ou de l'exactitude de la description, un brevet d'invention de quinze années, qui ont commencé à courir le 11 Août 1877, pour un perfectionnement du procédé connu sous le nom de Frézonnage, particulièrement applicable aux tissus petits-teints ou garance.

Article deuxième.

Le présent arrêté, qui constitue le brevet d'invention, est délivré à la Dme Vve Frézon, Sr Macé, Dme Macé, Sr Frézon, Picard et Dme Picard pour leur servir de titre.

A cet arrêté demeurera joint un des doubles de la description déposés à l'appui de la demande.

Paris, le 3 novembre mil huit cent soixante-dix-sept.

Pour le Ministre et par délégation :

Le Directeur du Commerce intérieur.

Figure 10: Example of a patent application document cover. This is patent application with archive code 1BB119863 from year 1877

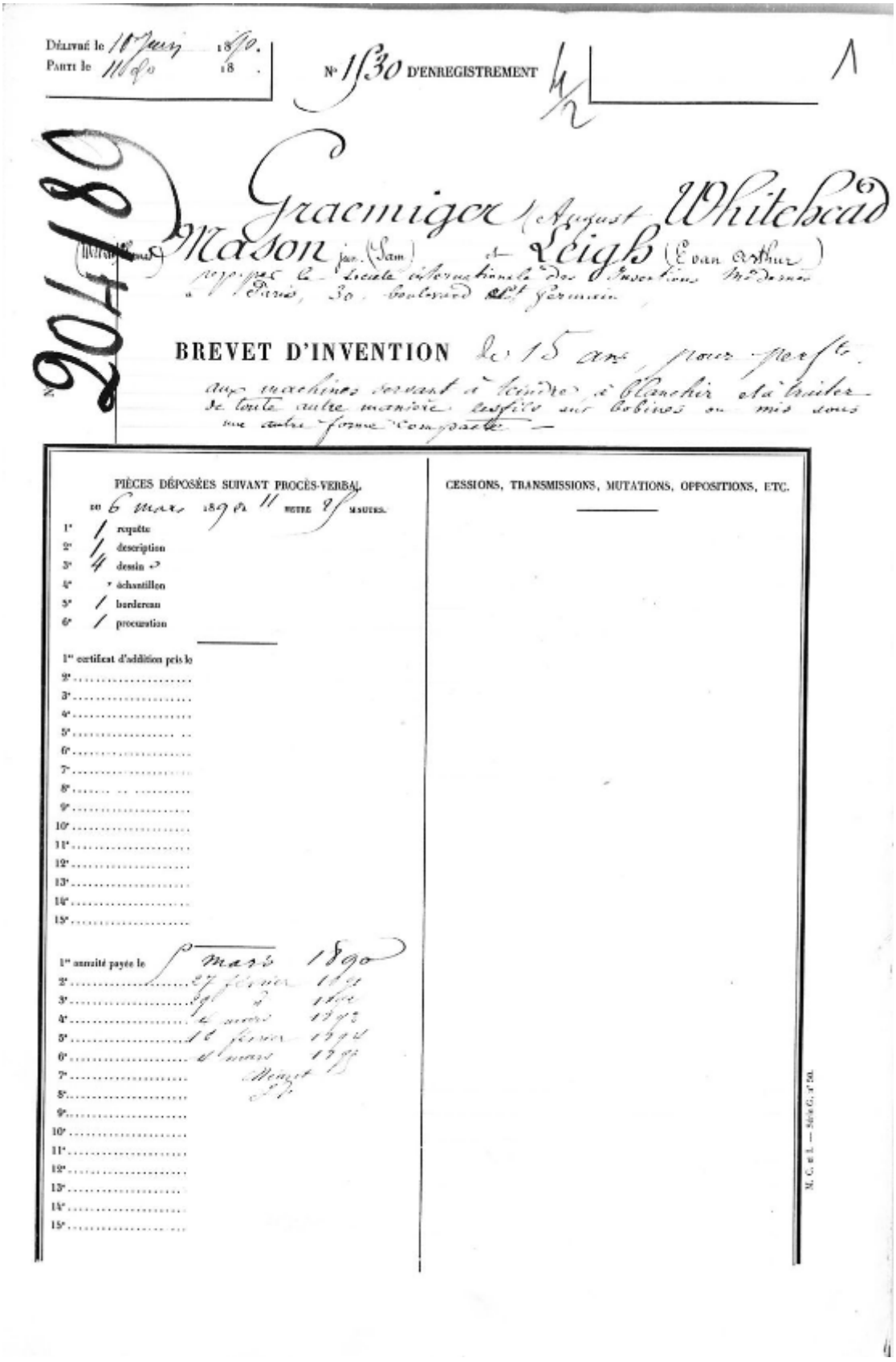


Figure 11: Example of a patent application document cover. This is patent application with archive code 1BB204189 from year 1890

D Novelty and Influence Measures

D.1 Notable Patents from Nineteenth-Century France

Table 15: Table of Notable French Patents (1791–1900)

Year	Inventor(s)	Title (short)	Reason	Sector	Application code
1791	Nicolas Leblanc	Large-scale soda-ash process	Birth of modern alkali industry	Chemistry	1BA12
1797	Firmin Didot	Stereotype printing process	Mass-edition printing	Industrial Arts	1BA2005
1800	Abraham-Louis Breguet	Tourbillon timekeeping regulator	Landmark in precision horology	Precision Instruments	1BA1792
1800	Joseph-Marie Jacquard	Programmable loom (punch-cards)	Automates patterned weaving	Textiles	1BA134
1806	Nicéphore & Claude Niépce	"Pyreolophore" heat-engine	Early internal-combustion engine	Machinery	1BA395
1820	Charles-Xavier Thomas de Colmar	Arithmometer calculator	First successful calculating machine	Precision Instruments	1BA1447
1830	Barthélemy Thimonnier (w/ A. Ferrand)	Chain-stitch sewing machine	First practical French sewing machine	Textiles	1BA3587
1832	Benoît Fourneyron	Water turbine	First efficient reaction turbine	Machinery	1BA3972
1839	Stanislas Sorel	Hot-dip galvanizing of iron	Corrosion protection by zinc coating	Mining & Metallurgy	1BA8634
1846	Adolphe Sax	Saxophone family	New woodwind instrument	Industrial Arts	1BB3226
1849	Charles-Xavier Thomas de Colmar	Industrial arithmometer	Mass-produced calculator	Precision Instruments	1BB8282
1851	Henri Giffard	Steam-powered airship (dirigible)	First powered dirigible concept	Marine & Navigation	1BB12226
1854	André-Adolphe Disdéri	Carte-de-visite photography	Made portrait photography mass-market	Industrial Arts	1BB21502
1857	Édouard-Léon Scott de Martinville	Phonautograph (sound writing)	First device to record sound	Industrial Arts	1BB31470
1858	Henri Giffard	Self-acting steam injector	Simple boiler feed device	Machinery	1BB36512
1859	Prudent-René Dagron	"Microscope-jewel" stereoscope	Microphotography pioneer	Precision Instruments	1BB41361
1859	Édouard-Philippe Carré	Absorption refrigerator / ice-maker	Early artificial refrigeration	Lighting & Heating	1BB41958
1860	Étienne Lenoir	Gas-fired internal-combustion engine	First commercially used IC engine	Machinery	1BB43624
1862	Alphonse Beau de Rochas	Four-stroke engine cycle	Principle of modern Otto cycle	Lighting & Heating	1BB52593
1865	Louis Pasteur	Wine-preservation by heating	Pasteurization applied to wine	Food Processing	1BB67006

Year	Inventor(s)	Title (short)	Reason	Sector	Application code
1866	Hippolyte Marinoni	Mechanical perfecting press	High-throughput newspaper printing	Industrial Arts	1BB71339
1867	Hippolyte Marinoni	Six-feeder cylinder press	Faster multi-feeder rotary press	Industrial Arts	1BB76392
1867	Joseph Monier	Iron-and-cement flowerpots	Origin of reinforced concrete	Agriculture	1BB77165
1868	Louis Ducos du Hauron	Color photography methods	First practical color processes	Industrial Arts	1BB83061
1869	Hippolyte Mège-Mouriès	Oleomargarine process	Invention of margarine	Chemistry	1BB86480
1871	Zénobe Gramme	Ring-armature dynamo	Breakthrough DC generator	Precision Instruments	1BB87938
1874	Émile Baudot	Rapid telegraph system	Baudot code & multiplexing	Precision Instruments	1BB103898
1876	Paul Jablochkoff	Arc “candle” lamp	City-scale electric lighting	Precision Instruments	1BB112024
1878	Auguste de Méritens	Dynamo-electric induction machine	Widely used arc-lighting dynamo	Precision Instruments	1BB123766
1879	Léon & Henri Serpollet; Claudius Richard	Instantaneous (flash) steam generator	Foundation of steam cars/boats	Machinery	1BB133362
1881	Auguste de Méritens	Electric autogenous welding	Early arc-welding process	Mining & Metallurgy	1BB146010
1882	Lucien Gaulard & John-Dixon Gibbs	AC distribution by transformers	Step-down/step-up power networks	Precision Instruments	1BB151458
1886	Paul Héroult	Electrolytic aluminum smelting	Hall-Héroult breakthrough	Mining & Metallurgy	1BB175711
1890	Clément Ader	Fixed-wing “avion”	Early airplane concept	Marine & Navigation	1BB205155
1891	Michelin & Cie (André & Édouard Michelin)	Detachable pneumatic tyre	Quick-change tire for cycles/cars	Road Transport	1BB214256
1892	François Hennebique	Reinforced-concrete system	Modern RC frames & beams	Construction	1BB223546
1895	Auguste & Louis Lumière	Cinématographe camera-projector	Birth of motion-picture projection	Industrial Arts	1BB245032
1895	Panhard & Levassor	Motor-car improvements	Standardized “Système Panhard”	Road Transport	1BB245276
1895	Louis-Michel Bullier	Practical acetylene for lighting/heating	Makes acetylene usable at scale	Lighting & Heating	1BB246000
1896	Georges Claude & Albert Hess	Acetylene storage system	Safe storage (toward dissolved gas)	Chemistry	1BB257679
1897	Panhard & Levassor	Gearbox with change-speed	Sliding-gear transmission	Machinery	1BB273375

D.2 Novelty validation: The case of the Aérostat patents

To-do: add a section showing how the Novelty measure predicts initial patents solving new problems. The case of Aérostat patents shows this clearly as the field figured out flying, then direction then structural forms and then propulsion. The initial patents, each time, hold higher novelty values.

D.3 stability of thresholds and the effect of CSLS across decades

To-do: add a figure to show this.

E Event Study Robustness

E.1 Productivity

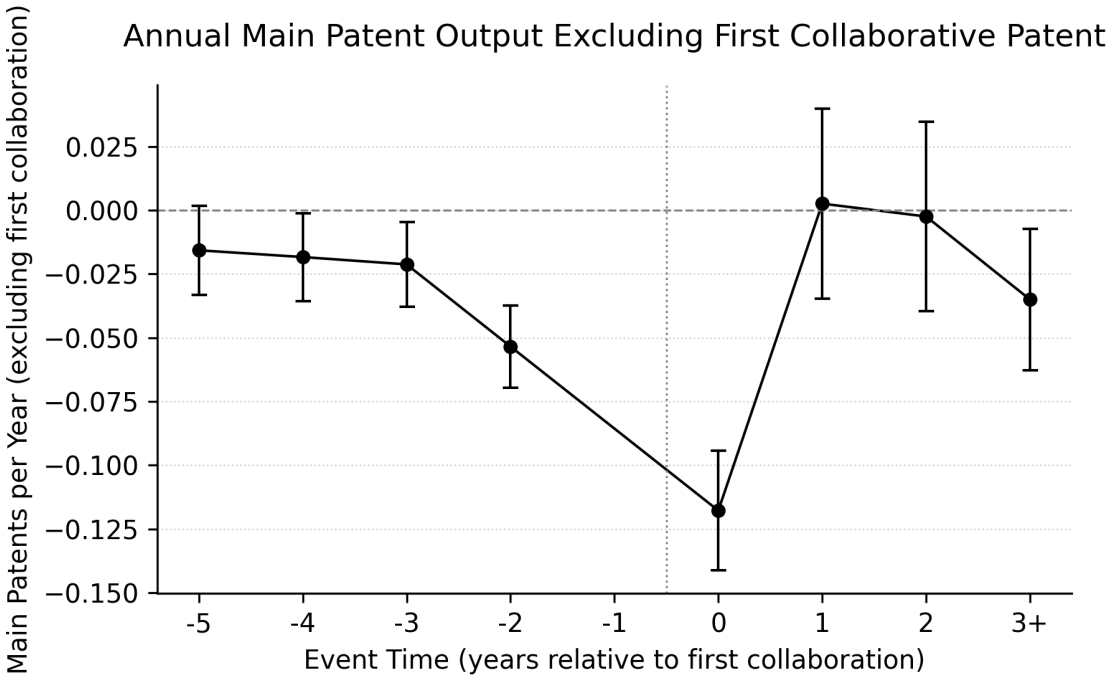


Figure 12: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is MainCount_excl_first_it, constructed identically to Figure 1 except that for k=0, the count excludes the first collaborative patent itself (i.e., MainCount_it - 1 at k=0 only). This robustness check isolates whether the k=0 spike reflects a mechanical inclusion of the collaborative patent or a genuine productivity increase. The negative k=0 coefficient confirms the spike in Figure 1 is mechanical. Sample, fixed effects, and clustering as in Figure 1. N=3,568,668 inventor-year observations.

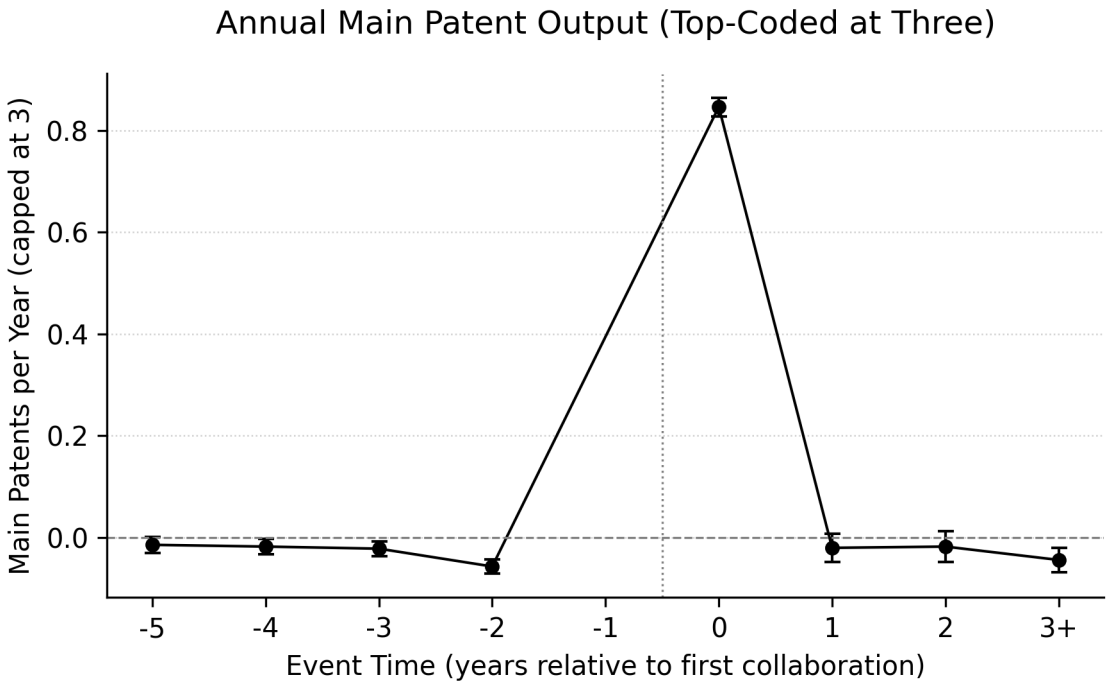


Figure 13: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is main patent count top-coded at three patents per year to reduce the influence of extreme outliers. Event time k=0 marks first collaboration; k=-1 is omitted. The k=0 coefficient (0.85, s.e. 0.01) is nearly identical to Figure 1, confirming that outliers do not drive the event-year spike. Pre-trends remain rejected. Sample, fixed effects, and clustering as in Figure 1. N=3,568,668 inventor-year observations.

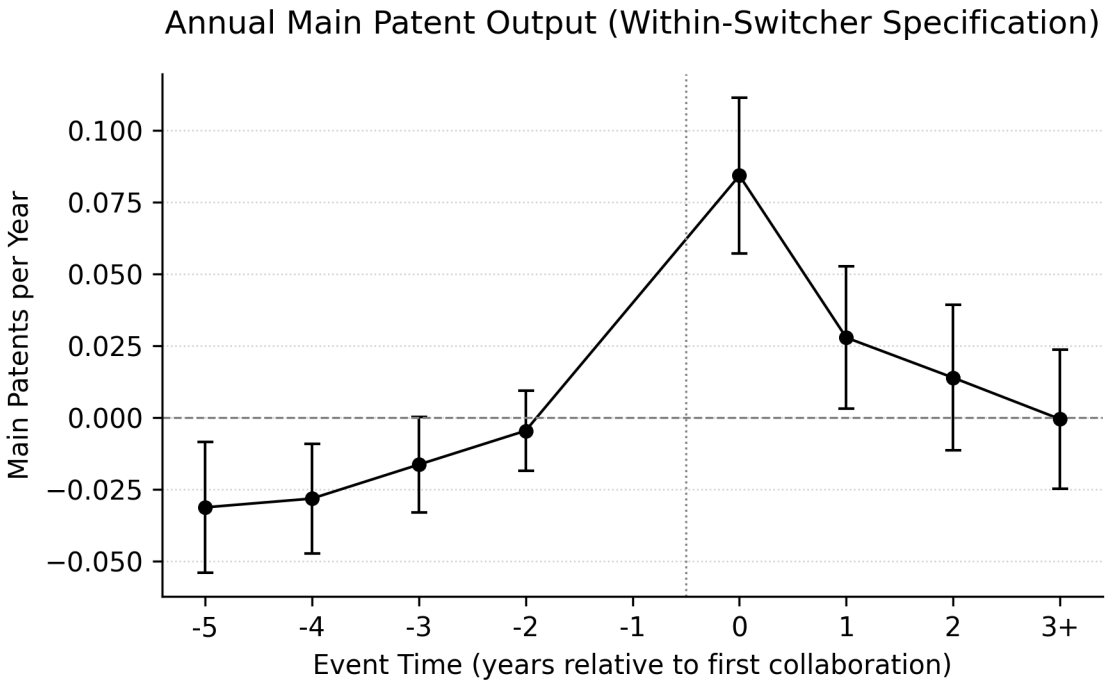


Figure 14: Point estimates and 95% confidence intervals from an event-study design estimated only on switchers (inventors who filed at least one solo patent before collaborating). The outcome is main patent count. Event time k=0 marks first collaboration; k=-1 is omitted. The specification absorbs inventor fixed effects and contemporaneous technology-class-by-year fixed effects. Event times k 3 are binned. Standard errors are clustered at the inventor level. This more conservative specification fails the parallel-trends test (F=2.8, p=0.025) and delivers a much smaller k=0 effect (0.08, s.e. 0.01) than the stacked design in Figure 1, consistent with the interpretation that most of the Figure 1 spike is mechanical. Sample: switchers only, 1791–1900. N=34,170 inventor-year observations.

E.2 Novelty

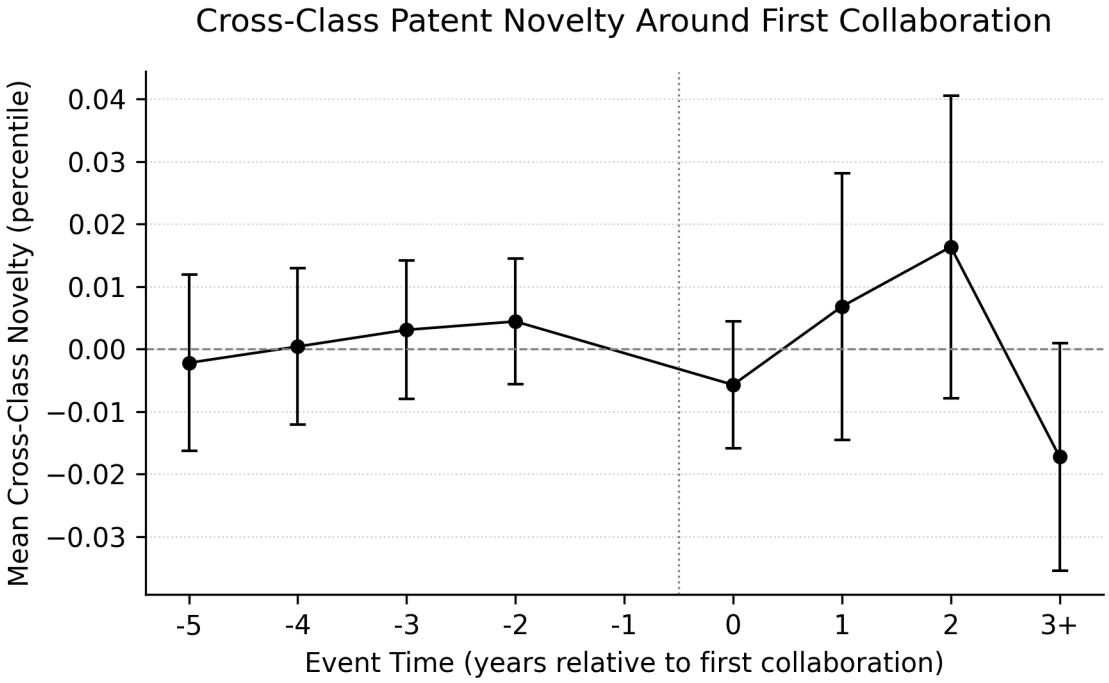


Figure 15: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is mean novelty across main patents filed in year t, restricted to patents in technology classes different from the inventor's pre-event modal class (cross-class novelty). Novelty is measured as the inverse percentile rank of backward semantic adoption: higher values indicate greater distinctiveness from prior art. Event time k=0 marks first collaboration; k=-1 is omitted. The sample includes switchers and never-treated controls, 1791–1900. The specification absorbs inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects. Event times k 3 are binned. Standard errors are clustered at the inventor level. Coefficients are small and statistically indistinguishable from zero, indicating no systematic change in cross-class patent originality following collaboration. N=1,057,035 inventor-year observations.

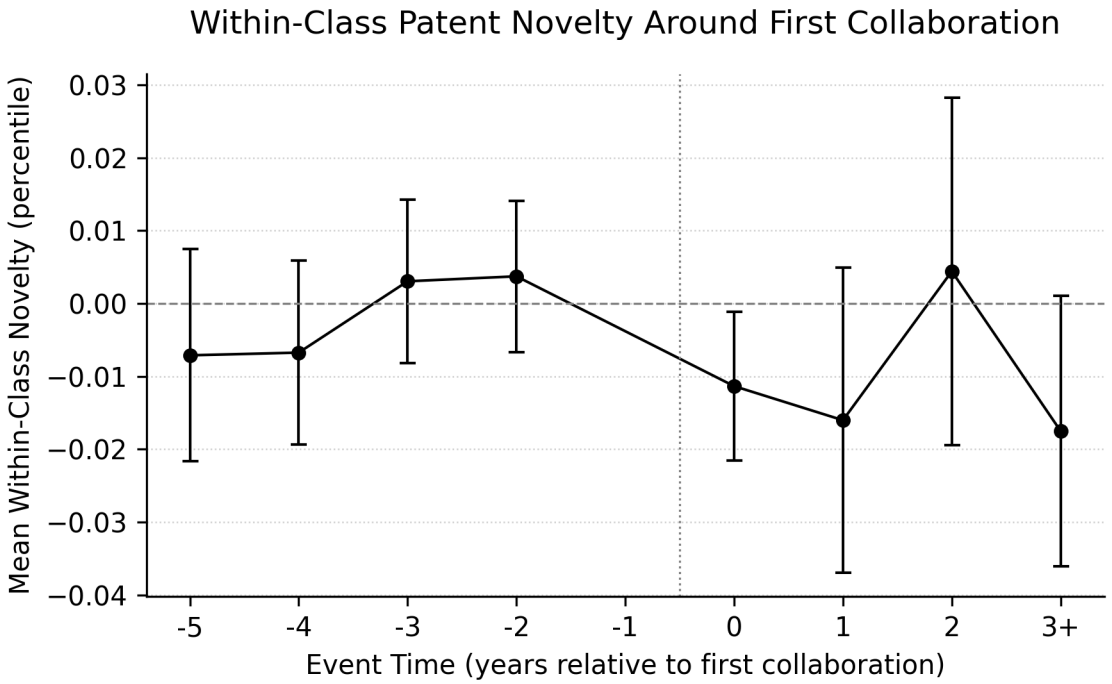


Figure 16: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design. The outcome is mean novelty across main patents filed in year t, restricted to patents in the same technology class as the inventor's pre-event modal class (within-class novelty). Novelty is measured as the inverse percentile rank of backward semantic adoption. Event time k=0 marks first collaboration; k=-1 is omitted. The sample includes switchers and never-treated controls, 1791–1900. The specification absorbs inventor fixed effects and pre-event-modal-technology-class-by-year fixed effects. Event times k 3 are binned. Standard errors are clustered at the inventor level. Coefficients are small and mostly insignificant, with a small negative effect at k=0 (-0.01, s.e. 0.01), confirming no systematic increase in within-class patent originality following collaboration. N=1,057,035 inventor-year observations.

E.3 Influence

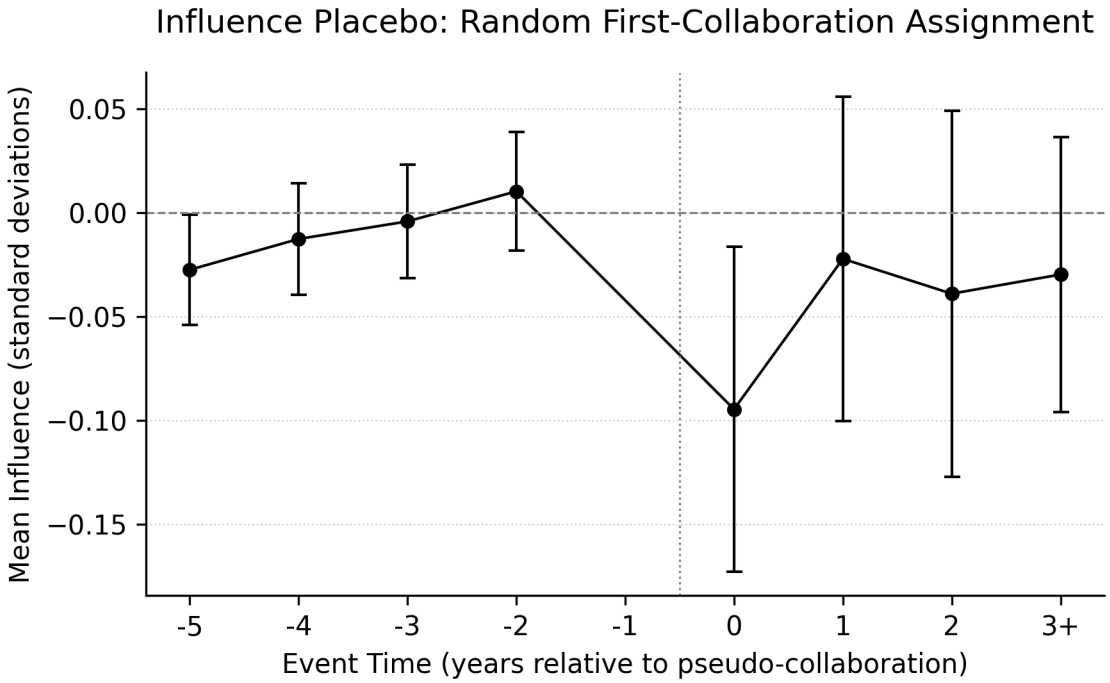


Figure 17: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design using a placebo treatment. For each always-solo inventor (who never collaborates), I randomly assign a pseudo-first-collaboration year drawn from the empirical distribution of true collaboration years among switchers, 1791–1900. The outcome and specification are identical to Figure 5. The flat path confirms that the influence pattern in Figure 5 is not spurious. Pre-trends are clean (F=1.5, p=0.21). N=403,914 inventor-year observations (always-solo sample).

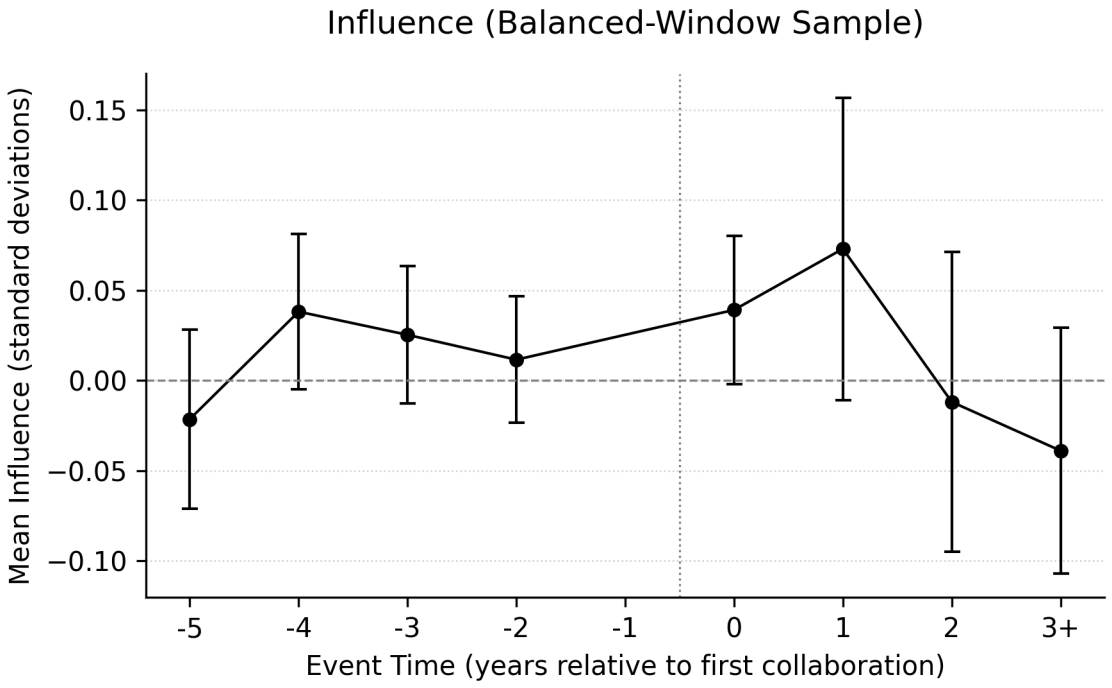


Figure 18: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design on a balanced sample. The sample includes only switchers who have at least one solo main patent in k [-5,-1] and at least one collaborative main patent in k [0,5], ensuring observable activity on both sides of first collaboration. The outcome and specification are identical to Figure 2. Results are nearly identical: k=0 effect of 0.04 SD (s.e. 0.02), peak at k=1 of 0.07 SD (s.e. 0.04). Pre-trends are not rejected (F=1.6, p=0.18). N=1,024,035 inventor-year observations.

E.4 Cross-class diversification

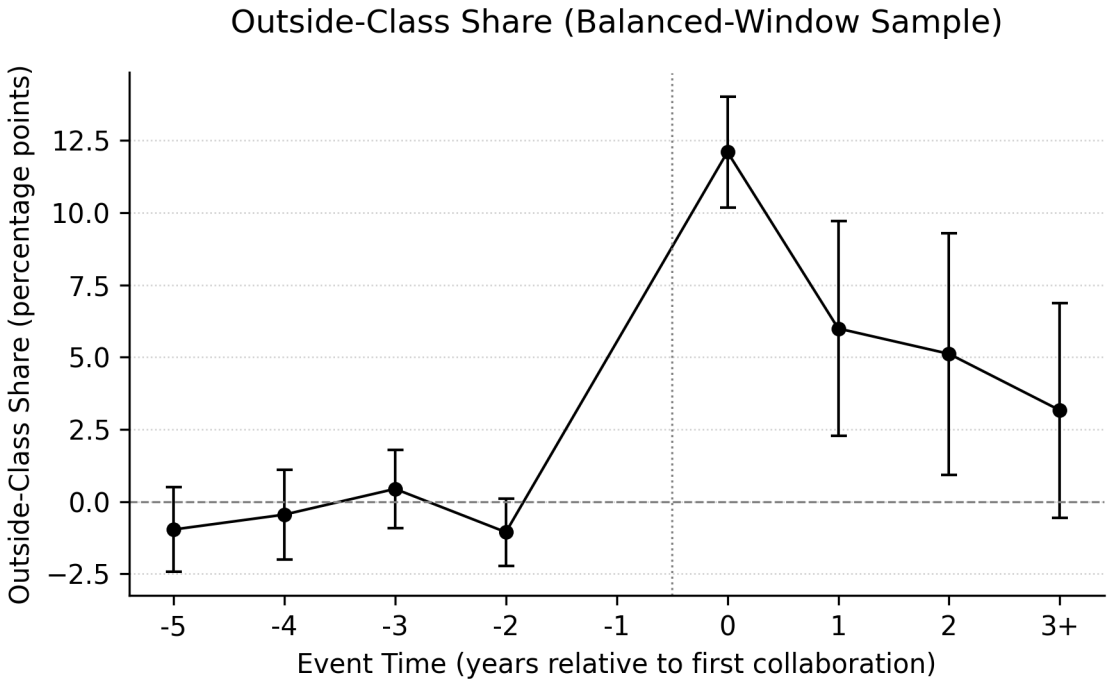


Figure 19: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design on a balanced sample. The sample includes only switchers who have at least one solo main patent in k [-5,-1] and at least one collaborative main patent in k [0,5]. The outcome, detrending procedure, and specification are identical to Figure 6. Results are nearly identical: k=0 jump of 12.1 pp (s.e. 1.0), k=1 effect of 6.0 pp (s.e. 1.9), k=2 effect of 5.1 pp (s.e. 2.1). Pre-trends are not rejected (F=1.5, p=0.20). N=65,473 inventor-year observations.

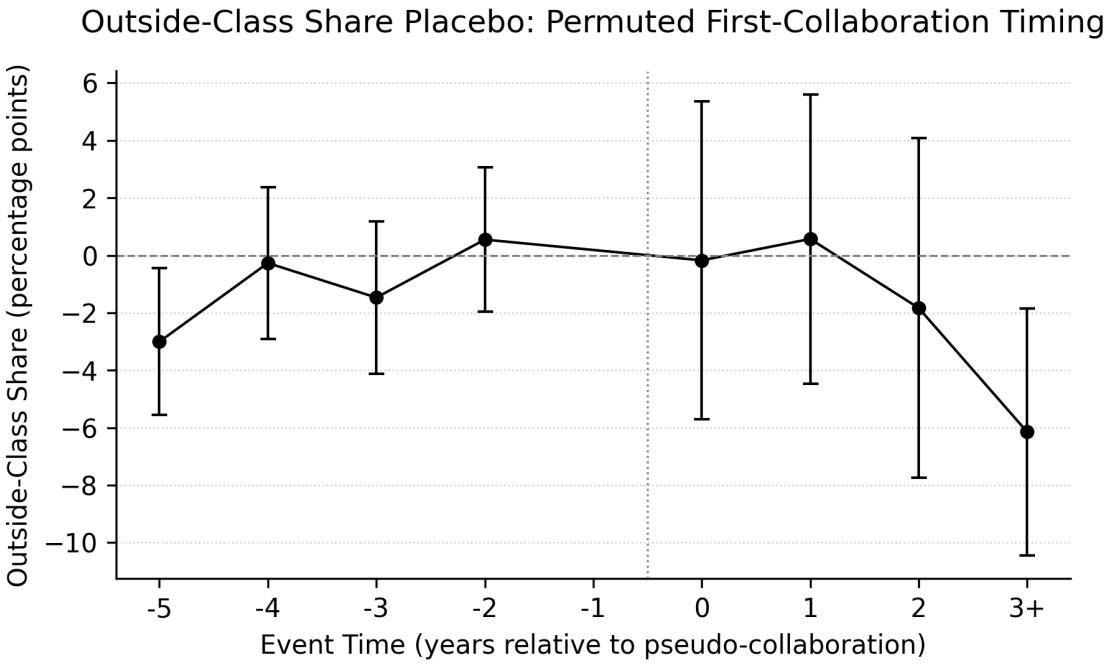


Figure 20: Point estimates and 95% confidence intervals from a Sun–Abraham stacked event-study design using permuted treatment timing. For each switcher, I randomly reassign a pseudo-first-collaboration year drawn from other switchers, breaking the link between true collaboration timing and outcomes while preserving sample composition. The outcome, detrending procedure, and specification are identical to Figure 6. The flat path (k=0 effect of -0.2 pp, s.e. 2.8, statistically insignificant) confirms that the diversification result in Figure 6 is not an artifact of specification or composition. Pre-trends are clean (F=1.9, p=0.11). N=121,261 inventor-year observations.

Table 16: Probit marginal effects

	Outcome: Collaborative Patent				
	Pre-reform	Post-reform	Full period	Full period	Full period
	(1)	(2)	(3)	(4)	(5)
15-years patent	-0.006	0.037***			
	(0.008)	(0.006)			
No. of Additions	-0.006*	-0.012***	-0.011***	-0.011***	-0.011***
	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
High Tech	0.029***	0.016***	0.017***	0.018***	0.018***
	(0.007)	(0.001)	(0.001)	(0.001)	(0.001)
Post 1844-reform			-0.023***	-0.019***	-0.018***
			(0.003)	(0.003)	(0.003)
Novelty (%)			-0.030***	-0.035***	-0.035***
			(0.002)	(0.002)	(0.002)
Influence				0.004***	0.004***
				(0.001)	(0.001)
Female-linked			0.047***	0.045***	0.045***
			(0.004)	(0.004)	(0.004)
Firm				-0.084***	-0.084***
				(0.003)	(0.003)
Foreign Import					0.012***
					(0.003)
Observations	12573	295800	307463	306997	306997
Pseudo R^2	0.002	0.001	0.003	0.007	0.008
Note:					*p<0.1; **p<0.05; ***p<0.01

F Patent level robustness

F.1 How different are collaboration patents?

Inventors in France could choose patent terms of 5, 10, or 15 years. Before 1844, the choice came with steep upfront costs: 300 francs for five years, 800 for ten, and 1,500 for fifteen. Over half the patents were of the shortest duration, and a fifth of the longest duration (see Appendix F.2). One might expect that longer patents, requiring greater investment, would be associated with joint ventures. But the probit model, Table 16, shows that a 15-year patent was no more likely to be collaborative than a shorter one.

The 1844 reform changed the system. Upfront fees were replaced by annual renewals of 100 francs. Suddenly, applying for 15 years became costless, an option that inventors could later abandon by simply ceasing payment. In this new regime, the data show a positive association between long patents and collaboration. But the effect is mechanical. After the reform, the 15-year patent became the default, so the variable no longer reflects inventors' choices.[22]

Collaborative patents are consistently less likely to generate the smaller improvement, or additions patents, (certificat de perfectionnement). Before the reform, collaborations were -0.62 percentage points less likely to produce follow-on additions patents; afterwards, -1.17 points less likely. The effect is persistent across different sample splits and with additional controls. Collaboration appears to deliver a, perhaps, more complete main invention, while solo inventors are more prone to pursue incremental extensions. Perhaps it was also harder to reassemble the team for smaller improvements.

High-tech[23] patents are far more likely to be collaborative. The effect is large: 2.85 percentage points in the pre-reform period, 1.59 points after. This suggests that collaboration mattered most

[22]The better measure is the patent's lapse date, which would give its actual duration. That information is not in my extracted data.

[23]In technology sectors characterized by rapid rates of technical progress. Such as textiles, machinery, mining and marine and navigation.

Table 17: Descriptive table for the probit marginal effects

		Missing	Overall	Pre-reform	Post-reform	P-Value
n			308373	12573	295800	
Collaborative Patent, n (%), n (%)	0		265534 (86.1)	10519 (83.7)	255015 (86.2)	<0.001
	1		42839 (13.9)	2054 (16.3)	40785 (13.8)	
15-years patent, n (%), n (%)	0		14469 (4.7)	9960 (79.2)	4509 (1.5)	<0.001
	1		293904 (95.3)	2613 (20.8)	291291 (98.5)	
10-years patent, n (%), n (%)	0		302038 (97.9)	9155 (72.8)	292883 (99.0)	<0.001
	1		6335 (2.1)	3418 (27.2)	2917 (1.0)	
5-years patent, n (%), n (%)	0		300239 (97.4)	6031 (48.0)	294208 (99.5)	<0.001
	1		8134 (2.6)	6542 (52.0)	1592 (0.5)	
High Tech, n (%), n (%)	0		218821 (71.0)	8442 (67.1)	210379 (71.1)	<0.001
	1		89552 (29.0)	4131 (32.9)	85421 (28.9)	
Female-linked, n (%), n (%)	0		302958 (98.2)	12380 (98.5)	290578 (98.2)	0.059
	1		5415 (1.8)	193 (1.5)	5222 (1.8)	
Firm-linked, n (%), n (%)	0		281388 (91.2)	12119 (96.4)	269269 (91.0)	<0.001
	1		26985 (8.8)	454 (3.6)	26531 (9.0)	
No. of Additions, mean (SD), mean (SD)		0	0.2 (0.7)	0.4 (1.0)	0.2 (0.7)	<0.001
Novelty (%), mean (SD), mean (SD)		910	0.5 (0.3)	0.5 (0.3)	0.5 (0.3)	<0.001
Influence, mean (SD), mean (SD)		1370	0.1 (0.9)	0.2 (1.0)	0.1 (0.9)	<0.001
Foreign Import, mean (SD), mean (SD)		0	0.0 (0.2)	0.1 (0.4)	0.0 (0.2)	<0.001

in sectors with rapid technical progress.

Novelty cuts one way and influence the other. In the full-period probit with baseline controls (Table 16, columns 3–5), moving a patent one percentile up the novelty distribution makes collaboration less likely by -0.034 to -0.030 percentage points. By contrast, patents whose language is taken up more by subsequent inventors (influence) are slightly more likely to be collaborative: a one-unit rise raises the probability of joint invention by 0.37 to 0.39 percentage points (using Models 4 and 5). Put plainly, highly original ideas tend to be pursued solo, while ideas that later diffuse and become templates for others are a bit more likely to originate in teams.

Finally, collaborative patents are more likely to be female-linked, less likely to be linked to a firm, and more likely when the patent involves a foreigner.

F.2 Probit Descriptive Statistics