The Labor Market Origins of Inventors

A New Inventor–LinkedIn Linkage Methodology

Elif N. Guler, Noah Stoffman, Wilson Wallace

We propose a new, large-scale linkage between U.S. patent inventors and LinkedIn professional profiles, built to support research on the career origins of inventors focusing on the relationship between education, employment histories, and innovative activity. The linkage covers more than 450,000 U.S. inventors. It connects USPTO patent records to self-reported LinkedIn education and employment histories using a multi-stage matching procedure that combines transparent, interpretable scoring rules with a supervised random-forest classifier trained on manually reviewed matches. This individual-level dataset connects inventors' complete educational and career trajectories, including institutions attended, fields of study, degree timing, employers, job titles, and employment spells, to their patenting outcomes. We built this infrastructure to study how labor market conditions at career entry shape who becomes an inventor, but the underlying linkage is a general-purpose research tool with applications well beyond our own project, and we plan to make a disclosure-protected version of it available to the broader research community.

In our presentation, we would focus primarily on the construction and validation of the linkage itself: the corporate affiliation crosswalk, the transparent scoring methodology, the random-forest classifier and its validation performance, and the tradeoffs involved in each design choice. We would briefly illustrate the value of the resulting data using summary statistics on coverage (match rates by cohort, technology class, and inventor characteristics) and a short illustration of the motivating labor-market-origins application.

Research Question

The constraint on research into the individual-level determinants of innovation is the absence of data that follow inventors over time. Patent records identify who invented what and for which assignee, but they say nothing about an inventor's educational background, the sequence of employers that preceded a given patent, or the career paths of similarly credentialed individuals who never patented at all. Administrative earnings and education records that could fill this gap are typically restricted, cover only a subset of the population, or cannot be linked to patent inventors at the individual level. This gap has made it difficult to answer basic questions about the inventor labor market: what determines who becomes an inventor, how career entry conditions shape the pipeline into inventive work, and how career trajectories differ between individuals who eventually patent and those who do not.

Our own motivating application is a study of how labor market conditions at labor market entry shape entry into invention. Tracing inventors from college through employment, graduate education, and inventive activity, we find that weak labor markets initially deter entry into invention, but that this effect reverses within several years, as recession-cohort entrants become

more likely to invent than their peers who entered the labor market in stronger years. This reversal is concentrated among individuals who enter graduate school shortly after college, and the inventors who emerge through this delayed path produce patents with different characteristics. Answering this question required building the underlying inventor–career linkage essentially from scratch, since no existing public dataset connects patent inventors to complete, dated career and education histories at this scale.

Data Sources

The linkage draws on three underlying data sources:

Patent data: We use the USPTO's Bulk Data Directory, specifically the PatentsView disambiguated inventor, location, and assignee datasets together with application and CPC technology classification files. Our baseline sample consists of 8,450,771 utility patents granted between January 1976 and September 2025. Records are merged on patent number and expanded to the patent–inventor–assignee triplet level, so that every observation carries the associated inventor identity, assignee, filing date, and the geographic locations of both inventor and assignee.

LinkedIn data: We use a 2022 snapshot of public LinkedIn profile data licensed from Nubela, which contains self-reported education histories (institutions, fields of study, and degree dates) and complete employment histories (employers, job titles and descriptions, employment locations, and spell start and end dates). These data allow us to reconstruct an individual's career trajectory over time, including periods before, during, and after any observed patenting activity.

CorpWatch corporate ownership data: We use CorpWatch's database of corporate hierarchies, built from SEC filings for U.S. public companies from 2003 through 2025, to resolve name changes, mergers, acquisitions, and parent–subsidiary relationships between patent assignees and LinkedIn-reported employers. We supplement CorpWatch's coverage with corporate relationships that we extract directly from free-text LinkedIn employment descriptions using a large language model, which allows us to capture smaller-scale acquisitions, rebranding, and spin-offs that fall outside CorpWatch's coverage of large public firms.

Matching Methodology

Linking patent inventors to LinkedIn profiles at the individual level is not a simple exact-match problem: names are reported inconsistently across the two sources, common names generate many false candidates, and employer names on LinkedIn frequently differ from the legal assignee name on a patent even when they refer to the same organization. Our matching procedure addresses these problems in four stages.

1. Candidate generation from inventor names: For each patent inventor, we generate a set of name variants. These name variants include adding or dropping middle names and middle initials, as well as replacing names with possible nicknames produced with the help of an LLM such as

Robert to Bob, Alexandar to Alex, or Michael to Mike. All the resulting name variants are used to find exact matches among LinkedIn profile names. This produces an initial set of candidate LinkedIn profile–patent pairs for each inventor.

2. Corporate affiliation crosswalks: Before matching individuals, we build a crosswalk that recognizes when a patent assignee and a LinkedIn-reported employer are related, even when the reported names differ. We link assignees to CorpWatch companies using an identifier-based crosswalk built through CRSP PERMCOs, Compustat CIKs, and CorpWatch identifiers, supplemented by standardized string matching for names that identifier-based linking misses.

We combine this with the LLM-extracted relationships from LinkedIn employment descriptions, using the number of employees reporting a given company and the number of independent descriptions corroborating a given affiliation to decide whether a relationship should be applied to all employees of a company or restricted to the individuals who reported it directly.

3. Transparent match scoring: Each candidate pair is scored using a composite measure built from four components: **(i)** a name score that rewards rarer names and more specific name agreement (e.g., a shared middle name); **(ii)** a company score based on IDF-weighted word overlap between the assignee and employer names, incorporating the corporate affiliation crosswalk; **(iii)** a timing score that rewards employment spells whose dates plausibly bracket the patent's filing date; and **(iv)** a location score based on haversine distance between geocoded inventor, assignee, and employment locations, with a string-based fallback when coordinates are unavailable. Since reported data is occasionally sparse in LinkedIn, the evidence available differs across observations; we normalize points earned by the maximum points attainable given the available evidence, yielding a score used to organize candidates by match quality and to construct a labeled training sample for the next stage.

4. Random-forest classification: We supplement the transparent scoring rule with a supervised random-forest classifier trained on high-scoring and manually reviewed positive and negative matches. In addition to the four score components, the classifier uses continuous string-similarity measures (Levenshtein ratio, token-set ratio, Jaccard similarity, overlap coefficient) between assignee and employer names, continuous geographic distances, raw employment-timing gaps, a job-title plausibility feature built from sentence-embedding similarity to reference lists of technical, ambiguous, and non-technical occupations, and missingness indicators.

The classifier is trained with cross-validated hyperparameter search optimizing AUC and applied to the full set of candidate pairs to produce a predicted match probability, which is our primary measure of match quality. Because a given inventor–patent observation can generate multiple candidate profiles, we compare all candidates within an observation to select the best-supported final match for identifying inventors.

This procedure differs from prior approaches to inventor-record linkage with its transparent matching algorithm. We combine four largely independent sources of evidence (which are name

rarity and specificity, employer identity including indirect corporate affiliation, employment timing, and geographical information) so that a match need not rely on any single noisy signal. Then, we use large language models not merely to generate features but to extend the coverage of a structured data source: relationships extracted from free-text LinkedIn descriptions let us recover small-scale acquisitions, spin-offs, and rebrandings that are absent from CorpWatch's SEC-based ownership records entirely. Lastly, rather than treating the interpretable scoring rule as the final product, we use it to construct a labeled training sample for a random-forest classifier that can learn nonlinear combinations of continuous features, including a semantic job-title plausibility measure that determines "innovative" job titles, and we validate this classifier's performance directly against manually reviewed matches, giving us a quantifiable precision/recall profile rather than an untested rule of thumb. Together, these choices let the linkage scale to the full USPTO grant universe while remaining auditable at every stage, from the underlying score components to the classifier's validation statistics.

Value to the Research Community

We see several distinct uses for this linkage beyond our own project.

- Labor economics of innovation: individual-level analyses of how education choices, graduate school timing, employer transitions, and career interruptions relate to inventive output.
- Studies of inventor mobility: because the data capture full employment spells rather than a single employer at the time of patenting, researchers can study job-to-job transitions among inventors, the relationship between mobility and patent value, and the diffusion of knowledge across firms via worker movement.
- Education and human capital research: the linkage connects detailed, dated degree and field-of-study information to a directly observed innovation outcome, which is otherwise difficult to obtain outside restricted administrative data.
- Comparisons between inventors and non-inventor peers: because LinkedIn coverage extends well beyond patent holders, the same infrastructure supports comparisons between inventors and observably similar individuals who did not patent, which is central to questions about selection into inventive careers.
- Methodological reuse: the corporate affiliation crosswalk (patent assignees, LinkedIn employers, and CorpWatch corporate hierarchies) and the LLM-assisted relationship-extraction and degree-classification pipelines are reusable components for other name- and entity-matching problems in innovation research, independent of the inventor application.

Plans for Public Dissemination

Because the underlying LinkedIn profile data are licensed for our own research use, our data agreement does not permit us to redistribute individual-level records. We therefore plan to disseminate the value of this linkage through freely available, aggregated and anonymized statistical tables rather than through release of a person-level crosswalk. Critically, this approach

lets us make the resource public without a restricted-access data use agreement, so any researcher can use it. We plan to construct and publicly release:

- Full technical documentation of the matching methodology and of the aggregation and suppression rules (including the scoring formulas, classifier feature set, training sample construction, and validation statistics) published as a standalone data appendix, together with validation benchmarks (precision/recall against our manually reviewed sample) so that other researchers can evaluate the reliability of the underlying linkage and repeat the matching processes.
- A school-by-cohort inventor "production" panel: for each institution attended by matched inventors, the number of inventors by graduating cohort and technology field/CPC class. Basically, the measure of how many inventors a given school has produced, and when.
- Age-at-invention profiles: the distribution of inventors' approximate age at first patent and across their patenting based on their bachelor graduation year.
- Graduate education attainment by school and field: the share of an institution's matched inventors who hold a graduate degree, broken out by degree type (Ph.D., master's, professional degree, etc.) and broad field of study, together with the distribution of years between undergraduate and graduate degree completion.

Patent-based sources such as PatentsView contain no information on inventors' educational institutions, degrees, or ages, while sources that do contain this information are individual, restricted-access administrative or survey data that cannot be linked to inventors at all. By converting our underlying individual-level linkage into open, disclosure-safe, school- and cohort-level statistics on education, age, and the pipeline into inventive careers, we can offer the I3 community a new public good without the licensing or access constraints that a person-level release would require.

We view the I3 Technical Working Group as an ideal venue to present this design and to get feedback on our transparent matching methodology enriched with random forest and which aggregate cuts of the data (e.g., additional geographic or firm-level breakdowns) would be most valuable to the community before we finalize the public release.