Cover note: The Commercialization Pipeline

Christopher Esposito (Georgia Tech) and Olav Sorenson (UCLA Anderson)

We propose to present *The Commercialization Pipeline*, a framework and open-source R package (`commpipe`) for estimating how any intervention affects the speed at which science moves toward commercial application. The framework traces knowledge trajectories along the shortest citation paths connecting scientific articles to downstream patents and estimates treatment effects on their commercial potential with a matched, stacked differences-in-differences design. The attached paper describes the data construction in full and demonstrates the method with two applications: transitions of research into regional hubs and into corporate settings.

Value to the research community. First, the framework reaches science that existing tools and freameworks miss. Much of the science that influencesr technology sits several steps upstream of the papers that patents cite directly (Ahmadpoor and Jones, 2017), but research on science-to-technology links has focused almost entirely on the final step. Our approach follows trajectories back to their earlier stages and asks what factors accelerate their transition from basic science to applied science. Second, most science is located upstream. Because many more papers sit several citation steps from a patent than at the paper-patent boundary, upstream interventions can affect many more papers. This yields large samples and ample statistical power, and it lets researchers study interventions that operate at scale. Third, the design is fully replicable. A researcher supplies only a list of treated papers, such as grant recipients, papers affected by a policy change, or publications from a particular institution. The package constructs multi-generation patent-to-paper trajectories, selects placebo trajectories, estimates stacked difference-in-difference regressions, and conducts diagnostics.

Presentation focus. We have two goals. First, we want to learn how the community would use the tool and how e can make it more valuable. Second, because the package automates core research tasks, we want to discuss openly what is being automated and whether researchers are comfortable with those choices. Our aim is for the package to conduct "mostly harmless econometrics": a transparent, conventional design whose every step can be inspected, rather than a black box. We would like feedback on which design decisions should stay fixed and which should be user-adjustable, and on which diagnostics the package should report by default. It currently reports coverage of the treated set, match rates, covariate balance, pre-trends, and placebo exposure to the treatment environment.

Dissemination plans. We will release `commpipe` on GitHub with documentation and a worked example that reproduces the paper's results. The network dataset we constructed, the shortest-path citation graph connecting papers to patents, will be deposited on Zenodo under a permissive license. The commercialization potential measure is distributed by its authors (Masclans, Hasan, and Cohen 2025) under a CC BY-NC license; to allow its authors to determine its distribution, our package directs users to download it from Masclan et al.'s repository rather than redistributing it directly.

The Commercialization Pipeline: A Framework for Identifying Factors that Accelerate Science to Technology

Christopher Esposito[1,*] Olav Sorenson[2]

September 22, 2026

Abstract

Despite widespread belief that scientific research fuels innovation and economic growth, the mechanisms that accelerate the translation of scientific discoveries into commercial applications remain poorly understood, particularly those not immediately connected to patenting. We develop an empirical framework for studying science-to-technology commercialization as a multi-step process. The framework traces knowledge trajectories along the shortest citation paths connecting scientific articles to downstream patents and uses a matched stacked differences-in-differences design to estimate how a given factor changes the speed with which these trajectories move toward commercial applications. We implement the framework in `commpipe`, an open-source R package that runs the full design for any user-supplied set of treated papers. To demonstrate its utility, we apply it to two factors: the science being pursued in a regional hub or in a corporate research setting, and show that hubs and corporate settings both accelerate progress toward commercialization.

[1]Jimmy and Rosalynn Carter School of Public Policy, Georgia Institute of Technology. [2]UCLA Anderson School of Management [*]Corresponding author: cesposito8@gatech.edu. The authors acknowledge research support from the UCLA Price Center for Entrepreneurship and Innovation.

1. Introduction

The private and public sectors spend billions of dollars on basic science. For example, in the United States, the National Science Foundation (NSF) funds roughly $10 billion in projects each year. The Department of Defense and the National Institutes of Health (NIH) combine for another $50 billion in annual spending. Private companies, from pharmaceutical manufacturers to quantum computing startups, also allocate tens of billions of dollars to scientific research (Arora et al. 2024). These investments pay off in the form of novel products and services that improve lives and that accelerate economic growth (Bush 1945; Cohen et al. 2002; Jaffe 1989; Mowery et al. 2001). If anything, science has been becoming increasingly important to innovation and economic growth (Fleming et al. 2019).

Although research strongly suggests that investments in science eventually yield economically valuable innovations, our ability to study these processes has been limited. Commercialization often requires a long pipeline from lab to product. Research to date has generally focused either on the correlations between scientific activity and invention, usually measured by patents or economic growth (Adams 1990; Jaffe 1989; Valero and Van Reenen 2019), or on the final step in the process, the research that immediately preceded a patent (Arora et al. 2024; Fleming and Sorenson 2004; Marx and Fuegi 2020; Siegel et al. 2003). But the transition from basic science to the commercialization of novel products and services can require years, if not decades, and often involves multiple steps to move from principles to applications (Ahmadpoor and Jones 2017; Van de Ven et al. 2008; Werth 1994). Existing approaches have had little insight into how anything up until that final step influences the pace of commercialization.

We propose an empirical framework for studying the factors that accelerate the advancement of science toward commercial applications along the full pipeline connecting basic science to applied science and on to commercialization. We first trace trajectories of scientific research as they progress from basic, fundamental discoveries to more applied scientific research and eventually to downstream commercial applications. We then use those trajectories to identify factors and

mechanisms that appear to accelerate their progression.

As an example, we use this framework to explore two factors that have been thought to accelerate the progression of science toward application: corporate engagement in science and commercialization hubs. Companies pursue scientific research to gain earlier access to academic research (Cohen and Levinthal 1990; Stern 2004). But they then probably steer their own research toward more commercial applications (Arora et al. 2024; Rosenberg 1990; Scharfmann et al. 2025). Geographic hubs, meanwhile, have been seen as important because they provide entire ecosystems to support the commercialization process, from early-stage funding to the mentoring of scientists (Clayton et al. 2018; Samila and Sorenson 2010; Stuart and Ding 2006).

To study how corporate engagement and embedding in hubs influence the translation of basic and applied science toward commercial applications, we trace knowledge trajectories connecting scientific articles to downstream patents through multi-step citation paths. Each citation step brings the trajectory closer to being cited by a patent. We trace these trajectories along the shortest citation path connecting a paper to a downstream patent, focusing our analysis on the most direct route from scientific research to commercial uses (Ahmadpoor and Jones 2017).

We characterize each paper according to its commercialization potential (CP), using the measures developed by Masclans et al. (2025). CP assesses the commercial potential of a scientific paper by comparing the textual similarity of its abstract to the abstracts of papers cited by patents. As the advances reported in papers come closer to being commercializable, CP rises. We therefore assess changes in the pace at which knowledge trajectories move toward becoming products through increases in this CP measure.

Our analysis follows a differences-in-differences logic. We identify trajectories that transition to hubs and to corporate settings and match them to "placebo" trajectories that did not undergo such transitions. We conduct these matches based on trajectories' pre-transition CP levels and trends, as well as their publication year and broad scientific field. We find that both the treated trajectories experience a post-treatment acceleration in their increases in commercialization potential, relative to the placebo trajectories that remain in academic settings and outside of hubs.

We offer two primary contributions. First and foremost, we propose a method for constructing and analyzing how the structure of the commercialization pipeline influences the speed at which it translates science into technology, and we release it as an open-source R package, `commpipe`, that any researcher can apply to their own set of treated papers. Second, we use the framework to examine how hubs and corporate science influence commercialization. Both appear to accelerate its movement toward application meaningfully.

We begin by introducing the datasets and describing the methods used to identify the shortest-citation-path knowledge trajectories connecting articles to downstream patents. We then detail the procedure used to match the transitioning trajectories to the counterfactual placebo trajectories. We then present two sets of results: the effect of transitions to hubs and corporations on CP, and the effect of such transitions on patenting hazards. We conclude with a discussion of the findings and possible areas for future research. Appendix C describes the `commpipe` package.

2. Data

Our publication data come from OpenAlex, the largest and most comprehensive repository of publication metadata. It covers more than 240 million publications from around the world. For our purposes, OpenAlex importantly includes information on publication year, research fields, the geographic location of authors, and paper-to-paper citations.

Our key dependent variable, the commercialization potential of the science reported in articles, draws on a dataset described in Masclans et al. (2025). They created their CP measure by training a machine learning model to predict the probability that an article would be cited by a patent based on the text of its abstract. Because OpenAlex only contains abstracts for a subset of its repository, using this measure requires us to limit our analysis to the 31 million papers with abstracts. In general, these papers have been published in higher-impact and more international journals than the full population of OpenAlex articles.

To build knowledge trajectories, we trace citations made by patents to scientific articles back multiple steps into the scientific literature. Beginning with patents means that we only include

science that has already become commercialized. That selection means that we probably miss trajectories with the longest paths as they end up being right censored in our data, not yet reaching a patent. Our approach therefore allows us to analyze what influences the speed to commercialization for science that becomes commercialized.

We build the trajectories backwards. We first identify scientific articles directly cited by USPTO patents using the "Reliance on Science" dataset (Marx and Fuegi 2020). We label these papers as "distance = 0 papers" because they sit right on the boundary between commercialization and science. If a paper receives citations from multiple patents, we randomly select and retain one of these citation edges.

We then identify the papers at distance = 1 from the nearest patent. Distance refers to the length of the citation path from a focal paper to a paper cited by a patent, so these papers all receive citations from distance = 0 papers. We repeat this process until we arrive at the distance = 10 papers, papers ten steps away from the nearest patent along the patent-paper citation graph.

Papers could fall on more than one path, potentially even being at different distances from being cited by a patent along the different trajectories. Following Ahmadpoor and Jones (2017), we restrict the citation graph to the edges that fall along the shortest paths, the geodesics, connecting papers to patents. For example, if a paper is cited by both a distance = 3 and a distance = 6 paper, we would retain only the citation from the distance = 3 paper. This restriction provides both conceptual and computational tractability. As the number of edges in a graph increases, the number of possible paths grows exponentially, to the point that computational analysis becomes infeasible.

This restriction to the shortest-path network also influences the interpretation of our results: because we analyze how hub and corporate environments increase commercialization potential along the shortest path trajectories, our estimates tell us how exposure to these environments make the shortest paths from papers to patents even shorter.

We connect these data to two sources to determine exposure to hub and corporate environments. We use the Duke DISCERN 2.0 dataset (Arora et al. 2023) to assign authors to corpo-

rate affiliations. We then geocode corporate and university affiliations using the full text address strings, often recorded at the lab or plant level, to identify the locations in which the science had likely been developed. In total, our dataset includes roughly 9.8 million scientific papers that fall on these shortest paths.

3. Methods

We adopt a differences-in-differences logic for our analysis, comparing trajectories that transition into hub locations or corporate environments to a set of otherwise-similar trajectories that do not undergo such transitions. To minimize the extent to which selection into these transitions might influence our estimates, we match the transitioning trajectories to the counterfactual placebo trajectories on multiple dimensions.

We match trajectories at both the paper and the trajectory level. We begin by identifying papers in which a trajectory first transitions to a hub location or a corporate environment. We identify corporate environments as those where the last author reports a corporate affiliation. In most fields, the principal investigator (PI), who oversees the lab or group producing the research, customarily appears last in the list of authors. We then define hub-produced papers as those for which the last author's affiliation has an address in one of the following U.S. cities: San Francisco, San Jose, Los Angeles, San Diego, New York City, Boston, Philadelphia (which includes Trenton/Princeton, NJ), Raleigh-Durham NC, Chicago, Houston TX, and Washington DC.

Let us label these transitioning papers as step $s = 0$ in trajectory i. A paper counts as a transition only if its immediate antecedent, $[i, s = -1]$, lies outside the relevant environment. We use the shortest-path citation graph to identify the immediate and two-step antecedent papers to these transitions (denoted by papers at the coordinates $[i, s = -1]$ and $[i, s = -2]$), as well as the papers that follow these transitions (denoted by $[i, s = 1]$ and $[i, s = 2]$).

We then match each transitioning paper to a paper in a trajectory that is in neither a hub nor a corporation in $s = -1$ or $s = 0$. From this set of possible matches, we first restrict the available candidates to those with $s = -2$ papers published in the same year and in the same broad field

($N = 26$) as the $s = -2$ paper for the transitioning paper, as well as in the same CP value decile. For each remaining candidate, we calculate the change in CP from $s = -2$ to $s = -1$ and choose the one closest to the CP change observed from $s = -2$ to $s = -1$ for the transitioning paper, matching without replacement. We discard matched pairs whose changes in CP differ by more than 0.05 (a caliper). In other words, we use nearest-neighbor matching on CP changes combined with coarsened-exact matching on field, year, and CP levels. Together, these restrictions should ensure that the transitioning trajectories and the placebos had been advancing toward commercialization at nearly the same rates prior to the transitions to hubs and commercial settings.

Because the matching procedure requires information on trajectories with at least three observable steps ($s = -2$ to $s = 0$), our 9.8 million papers yield far fewer usable trajectories: over half a million candidate placebo trajectories satisfy these requirements. This restriction tends to drop shorter trajectories (i.e., more applied areas of science).

We identify 39,498 papers that transition to hubs and 9,706 that transition to corporations in $s = 0$. Of these, 39,066 hub and 9,531 corporate transitions have complete pre-period trajectories. We match 38,256 (97.9%) of the hub-transitioning trajectories and 9,320 (97.8%) of the corporation-transitioning trajectories to placebo trajectories. Unmatched trajectories either had no candidate placebo in their stratum or exceeded the caliper.

Matching generated two datasets, described in Table 1. The first consists of trajectories that transition to hubs in $s = 0$, matched to placebo trajectories that did not. The second comprises trajectories that transition to corporations in $s = 0$, matched to placebos that did not. Table 1 describes each dataset and the balance of the transitioning and placebo papers in each. The matching yielded almost identical distributions of change-in-CP from $s = -2$ to $s = -1$, with mean residual within-pair differences of only 0.002 to 0.003. The high precision of these matches stems in large part from the large number of candidate placebos available after the coarsened exact matching on field, publication year, and CP level. Trajectories in the hubs panel had an average of 625 candidates, while those in the corporations panel had an average of 622 potential matches.

Table 1 also reports the shortest-path citation distances to the nearest patent for transitioning

Table 1: Balance of Data Panels

	Treatment: Transition to Hubs		**Treatment: Transition to Corporations**	
	Treated	**Placebo**	**Treated**	**Placebo**
Number of trajectories	38,256	38,256	9,320	9,320
Mean publication year $s = -2$	2007.81	2007.81	2007.46	2007.46
Modal publication field $s = -2$	Engineering	Engineering	Engineer	Engineer
Mean change in CP ($s = -2$ to $s = -1$)	0.055	0.055	0.074	0.074
Mean bin size used for matching	624.9	624.9	621.7	621.7
Residual change in CP after matching	0.002	0.002	0.003	0.003
Mean shortest distance to patent $s = -2$	4.64	4.61	3.95	4.23
Mean shortest distance to patent $s = -1$	3.64	3.61	2.95	3.23

and placebo trajectories. We would not necessarily expect balance on these dimensions. Most obviously, the matching procedure did not include these measures as criteria. More importantly, citation distance to the nearest patent provides a forward-looking measure that may itself depend on the transition. For example, a trajectory that shifts into a corporate setting might realize a patent at an earlier step than its matched trajectory that did not (leading to a shorter path length to the nearest patent, even in steps -2 and -1).

Rather than providing a diagnostic on the success of the matching procedure, distance to the nearest patent points to the possible causal effects of transitioning in accelerating commercialization. With each step, the shortest path distance to a patent mechanically falls by one. Because trajectories move closer to patenting with each step, level differences across the transitioning and placebo groups suggest a causal effect of the transition on path length. Transitions to corporations reduce the distance of otherwise-similar trajectories to the patent boundary by $(4.23 - 3.95) = 0.28$ citation steps. Transitions to hubs, by contrast, leave distance essentially unchanged (4.64 versus 4.61), foreshadowing the more modest effect of hubs on patenting that we report below.

The shortening of science-to-commercialization trajectories induced by the earlier patenting of transitioning trajectories also affects the composition of the sample. When a trajectory transitions to a patent, it ends. Naïve estimation that does not account for the changing sample could therefore understate the effect of transitioning on CP (if increases in CP correlate positively

with patenting, as expected).

Table 2 reports the number of observations with a CP value in transitioning and placebo trajectories relative to the matched transitioning papers. Both sets of trajectories have the same number of observations at steps -2 and -1, as required by the matching procedure. At step 0, a small share of transitioning papers (1.0% for hubs and 2.5% for corporations) lack a CP value. Larger differences in the number of observations emerge in steps 1 and 2, as trajectories exit the panel upon reaching a patent.

Table 2: Observations by Step and Treatment Group

	Treatment: Transition to Hubs		**Treatment: Transition to Corporations**	
Step	**Treated**	**Placebo**	**Treated**	**Placebo**
-2	38,256	38,256	9,320	9,320
-1	38,256	38,256	9,320	9,320
0	37,890	38,256	9,086	9,320
1	27,695	28,532	5,305	6,223
2	19,403	19,535	3,296	4,040

Finally, Table 3 reports the commercialization potential (CP) of paths by step. Prior to the transitions, transitioning and placebo trajectories have identical commercialization trajectories. Differences emerge in step 0, as trajectories that transition to hubs increase in CP by 0.04 more than their placebos and trajectories that transition to corporations by 0.09 more than their placebos. These differences attenuate in steps 1 and 2.

Table 3: CP by Step and Treatment Group

	Treatment: Transition to Hubs		**Treatment: Transition to Corporations**	
Step	**Treated**	**Placebo**	**Treated**	**Placebo**
-2	0.39	0.39	0.48	0.48
-1	0.44	0.44	0.55	0.55
0	0.55	0.51	0.67	0.58
1	0.55	0.53	0.66	0.59
2	0.57	0.56	0.67	0.62

4. Effects of Treatments on the Commercialization Potential of Science

We next estimate the change in commercialization potential (CP) in a linear model:

$$CP_{i,s} = \sum_{\substack{s=-2 \\ s\neq -1}}^{2} \beta_s^t \left(\text{StepsToTransition}_{i,s} \times \text{Treated}_i\right) + \alpha_i + \gamma_{p(i),s} + \varepsilon_{i,s}. \quad (1)$$

We estimate Equation 1 as a stacked regression (Cengiz et al. 2019). Conventional two-way fixed effects estimators can produce biased estimates when treatment effects vary across units and over time, because they implicitly use already-treated units as controls (Goodman-Bacon 2021). Our design avoids this problem. Each transitioning trajectory is aligned to its own transition at $s = 0$, and its matched placebo is untreated at $s = -1$ and $s = 0$, so each matched pair forms its own clean treated-versus-control comparison. The model includes trajectory fixed effects, α_i, and matched-pair $\times$ event-time fixed effects, $\gamma_{p(i),s}$, where $p(i)$ denotes the matched pair containing trajectory i. The pair-by-step fixed effects absorb the common commercialization profile across steps and restrict identifying variation to within-pair, within-step contrasts between a transitioning trajectory and its own placebo. No already-treated trajectory ever serves as a control, and comparisons across pairs are not used for identifying variation. The coefficients β_s^t therefore measure the differential change in commercialization potential for transitioning trajectories relative to their matched placebos, by step, relative to the reference step $s = -1$.

Table 4 Column 1 provides estimates for transitioning to a hub while Column 2 reports the estimates for transitioning to a corporate environment. As one would expect, commercialization potential rises with each step in both transitioning and placebo trajectories. Transitioning trajectories nevertheless experience even faster acceleration in this potential. Consider first the transition to hubs. While placebo trajectories increased in CP by 0.065 between the reference step ($s = -1$) and the next step ($s = 0$), transitioning trajectories increased in CP by 0.043 more. In other words, transitioning to a hub appeared to accelerate the movement toward commercialization by roughly 65%. This positive effect of transitioning to a hub persists for at least two additional steps after the

transition.

Table 4: Regression Estimates for Changes in Commercialization Potential

	(1) Transitions to Hubs	(2) Transitions to Corporations
$\beta^t_{s=-2}$	-0.0000521^{*}	-0.0000444
	(0.0000314)	(0.0000662)
$\beta^t_{s=0}$	0.0427^{***}	0.0884^{***}
	(0.00169)	(0.00335)
$\beta^t_{s=1}$	0.0193^{***}	0.0700^{***}
	(0.00243)	(0.00546)
$\beta^t_{s=2}$	0.00854^{***}	0.0512^{***}
	(0.00328)	(0.00759)
Trajectory FE	Yes	Yes
Pair $\times$ step FE	Yes	Yes
Observations	324,335	74,550
Trajectories	76,512	18,640

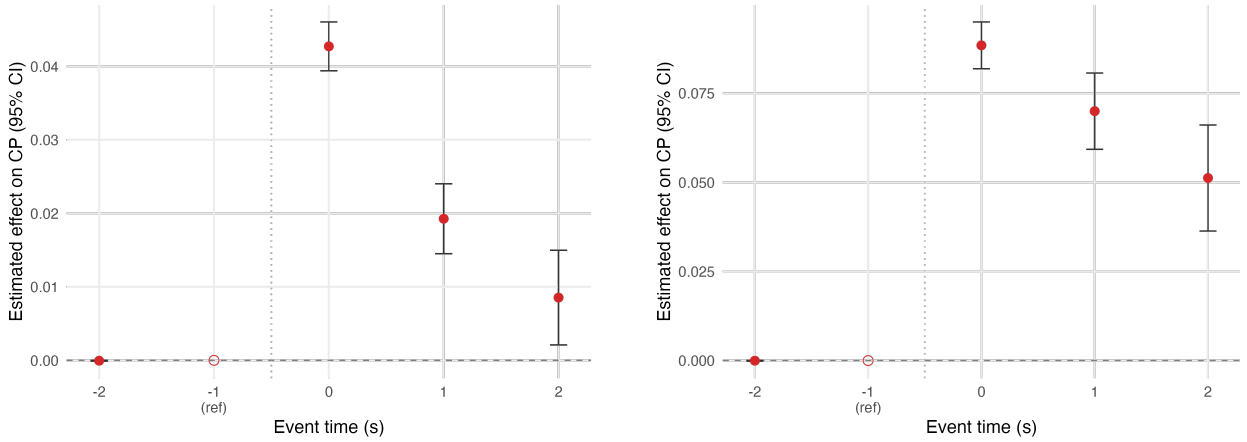
Note: The estimating sample stacks each transitioning trajectory with its matched placebo. Coefficients β^t_s give the differential change in commercialization potential for transitioning trajectories relative to matched placebos at each step, relative to the reference step $s = -1$. All models include trajectory and matched-pair $\times$ event-time fixed effects; the latter absorb the common across-trajectory step profile. Standard errors, clustered at the matched pair, in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The lagged coefficient in Table 4 Column 1, $\beta^t_{s=-2}$, also suggests that the transitioning sample contains minimal selection into hubs. The coefficient offers a test of the parallel pre-trend assumption, a necessary condition for being able to interpret the differences-in-differences estimate as being causal. The coefficient is marginally significant ($p = 0.10$), but its magnitude is negligible: it represents about 0.1% of the estimated $s = 0$ transition effect.

Column 2 of Table 4 reports results for transitioning into corporations. The coefficients tell a similar story as the transitions to hubs: although all trajectories increase in commercialization potential with each step, transitioning to a corporate setting accelerates that process. The magnitudes of the effect sizes for the corporate effect are roughly double those for hubs. The lagged coefficient is small and statistically indistinguishable from zero, again suggesting that selection into these transitions probably accounts for little of the estimated effect.

Figure 1 plots the coefficients associated with the StepsToTransition^t_i variables against

event time, with 95% confidence intervals. The figures show (1) the lack of substantive pre-trends, (2) the instantaneous impact of hubs and corporations on CP in $s = 0$, and (3) the persistence of the treatment effects in the subsequent steps, $s = 1$ and $s = 2$.



(a) Effect of Transitions to Hubs

(b) Effect of Transitions to Corporations

Figure 1: Event Plots

5. Effects of Transitions on Patenting Hazards

Although the commercialization potential measure gives us a continuous dependent variable that allows us to observe changes in the proximity to commercialization from one step to the next, we can also directly estimate the effect of transitioning to a hub or to a corporate environment on the observed rates of patenting for trajectories.

Examining patenting rates directly can also help us to address a potential methodological issue with our CP regressions. When a trajectory culminates in a patent, its sequence ends and no longer contributes to our coefficient estimates. Trajectory attrition changes the composition of the data set for $s = 1$ and $s = 2$. To the extent that transitions and patenting correlate positively with increases in CP, attrition could negatively bias our estimated effect of CP on later steps. In other words, it could explain the apparent attenuation of the transition effects in our prior analyses.

To estimate the effects of transitions to hubs and corporations on the rate of patenting, we estimate a discrete-time hazard model (as a pooled logit). The dependent variable is an indicator

equal to one if a trajectory patents at a given step and zero otherwise. We restrict the analysis to $s = 1$ and $s = 2$ since our construction of the trajectories means that they could never result in a patent before that. Because the model conditions on survival up to each step, removing trajectories from the risk set once they patent, it accounts for the right censoring induced by patenting.

Because each trajectory follows a shortest path, whether it patents within two steps depends largely on its distance to the nearest patent at the transition. The hazard model therefore captures timing: a positive treatment coefficient indicates that transitions move trajectories closer to the patent boundary, so that they patent at an earlier step, rather than a change in whether they eventually patent at all.

We again define a trajectory as transitioning to a hub or to a corporate setting based on whether these shifts occurred at the $s = 0$ paper. We include an indicator variable for each transition, an indicator variable for $s = 2$ (which one could see as a non-parametric estimate of the step "clock"), and the interaction between transitioning and $s = 2$. We cluster the standard errors at the matched-pair level.

Table 5 reports the results. Both types of transitions increase the patenting hazards post-transition, with the corporate effect roughly three times the hub effect. The negative interaction effect, however, suggests that this effect attenuates over time (and therefore also suggests that the attenuation in the CP models does not stem from attrition in the sample). These estimates depend on screening placebos out of both environments. If we instead screen placebos only on their own environment, allowing, for example, corporate papers into the placebo pool for the hub analysis, the treatment coefficients fall to 0.099 for hubs and 0.326 for corporations, consistent with placebos that pass through the other environment patenting sooner.

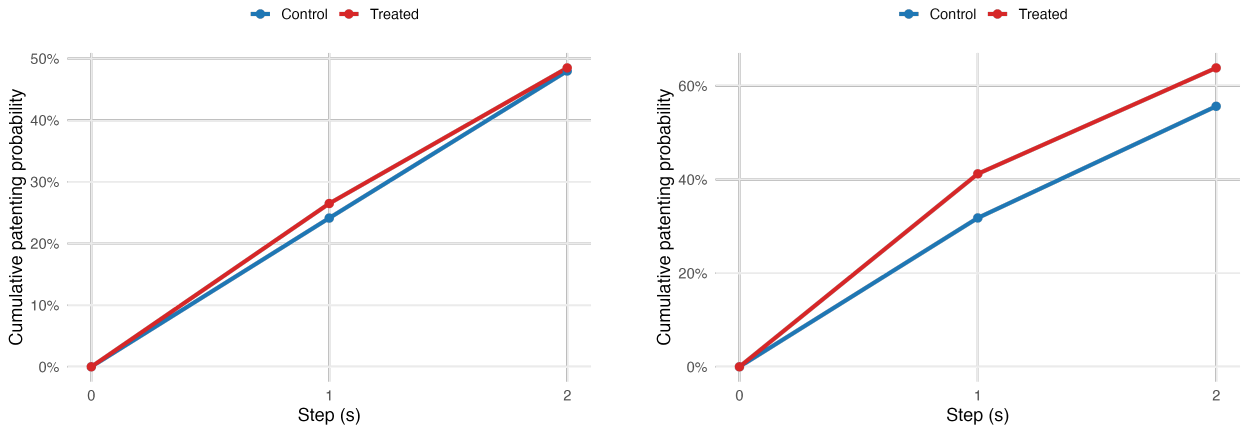
To facilitate interpretation of the hazard estimates, we convert the estimated step-specific patenting rates into cumulative patenting probabilities by step for transitioning and placebo (control) trajectories. Using the fitted model, we compute the probability that a trajectory patents by step one and by step two, accounting for the fact that patenting at earlier steps removes trajectories from the risk set at later steps. We plot these values in Figure 2.

Table 5: Effects of Transitions on Patenting Hazards

	(1) Effect of Hubs	(2) Effect of Corporations
Intercept	-1.146^{***}	-0.764^{***}
	(0.0119)	(0.0222)
$\beta_{s=2}$	0.366^{***}	0.143^{***}
	(0.0174)	(0.0345)
Treated	0.125^{***}	0.408^{***}
	(0.0144)	(0.0271)
Treated $\times \beta_{s=2}$	-0.197^{***}	-0.257^{***}
	(0.0240)	(0.0479)
Observations	133,654	30,471

Note: Discrete-time hazard estimated as a pooled logit on the at-risk panel at $s = 1$ and $s = 2$; trajectories leave the risk set once they patent. Standard errors, clustered at the matched pair, in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The resulting cumulative patenting curves reveal that transitions to hubs and to corporate settings accelerate patenting. Treated trajectories exhibit higher cumulative patenting probabilities immediately after treatment. By step one, 26.5% of hub-transitioning trajectories have patented, compared with 24.1% of their placebos; by step two, the two groups have converged (48.5% versus 48.0%). Hubs therefore appear to speed patenting rather than to raise its eventual likelihood. The corporate gap is larger and more persistent: 41.2% versus 31.8% by step one and 63.8% versus 55.6% by step two.



(a) Cumulative Patenting for Hubs

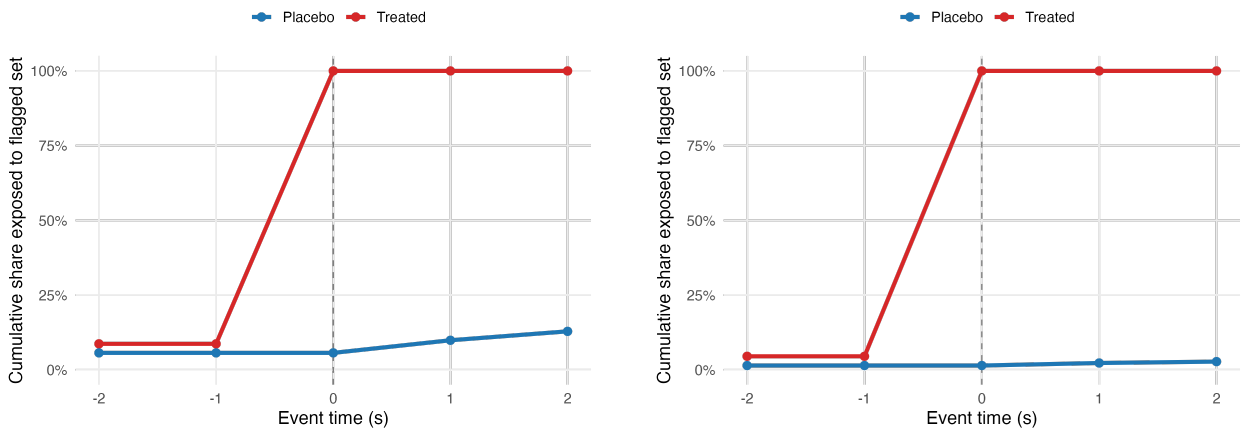
(b) Cumulative Patenting for Corporations

Figure 2: Cumulative Patenting Rates

6. Changing Counterfactual Composition

Our differences-in-differences approach examines the effects of transitions, to a hub or to a corporate setting, at step $s = 0$. Our analysis therefore estimates the effect of a transition rather than the absolute effect of being in a hub or a corporation. The distinction matters for two reasons. First, our data construction eliminates all trajectories that remain in hubs or in corporations from the beginning of the trajectory until commercialization. Second and more subtly, our placebo cases sometimes transition at a later point in time (i.e., $s = 1$ or $s = 2$). These transitions likely raise the CP and the patenting hazards of these trajectories, leading to measurement error in our "treatment" and a downward bias in our estimates of the effect of these transitions.

Figure 3 illustrates this issue. It plots the cumulative share of trajectories that have passed through each environment, by step, for transitioning and placebo trajectories. A small percentage of trajectories begin at $s = -2$ in the relevant environment (8.7% of hub-transitioning and 5.7% of placebo trajectories in hubs; 4.5% and 1.4% in corporations), but by construction, all transitioning trajectories enter the relevant environment at step $s = 0$, while no placebos occupy it at $s = -1$ or $s = 0$. However, the figure reveals that some placebo trajectories transition into hubs or corporations in steps $s = 1$ and $s = 2$: by step two, 12.9% of hub placebos and 2.7% of corporate placebos have ever been in the relevant environment.



(a) Cumulative Presence in Hubs

(b) Cumulative Presence in Corporations

Figure 3: Presence in Hubs and Corporations for the Treated and Control Groups

This changing composition in the counterfactual set over time may partially explain the attenuation across steps of the estimated transition effects on CP and patenting hazards. As placebo trajectories enter hubs and corporations at later steps, their outcomes begin to converge toward those of the earlier transitioning trajectories, reducing the estimated gap in subsequent periods. Attenuation therefore may reflect the expanding exposure of the placebos to hubs and corporations rather than a decay in the underlying commercialization advantages of these settings.

7. Discussion

We offer an empirical framework for extending the study of commercialization back from the boundary between papers and patents deep into the scientific pipeline leading up to those innovations. We do so by tracing scientific trajectories through citations, finding the shortest paths connecting science to patents.

The primary contribution of our research is to introduce a rigorous, scalable, and plausibly causal framework to study the factors that accelerate the full science-technology commercialization pipeline. Existing research focused on the subset of science that directly impacts technology (Arora et al. 2021; Bikard and Marx 2020; Marx and Fuegi 2020; Scharfmann et al. 2025), and thus does not address the many steps over the course of years that are taken by scientists to prepare scientific knowledge for commercialization. Our framework thus enables researchers to study the commercialization process upstream. Given the large volume of scientific effort that is expended to prepare knowledge for commercialization, this framework has the potential to substantially accelerate the commercialization process and improve the return on scientific investments.

As an example, we used our framework to estimate the importance of commercialization hubs and of corporate science on the speed of commercialization. Our results suggest that both have substantial effects, accelerating the speed with which ideas become commercial innovations.

Our framework can be readily deployed to study additional factors that may affect commercialization pipelines. For example, factors at the level of individuals, networks, institutions, and nations may all affect the speed of commercialization (Esposito 2026; Koffi and Marx 2026; Marx

and Hsu 2022). By combining shortest-path trajectories connecting science produced by such entities to downstream patents (Ahmadpoor and Jones 2017), combining the papers in those trajectories with information on their commercialization potential (Masclans et al. 2025), and by matching the "treated" trajectories to placebo controls, our framework can be extended to study multitudes of factors that may affect the commercialization pipeline. The `commpipe` package makes such extensions straightforward: a researcher supplies a list of treated papers, and the package constructs trajectories, matches placebos, and estimates both the CP event study and the patenting hazard.

References

Adams, J. (1990). Fundamental Stocks of Knowledge and Productivity Growth. *Journal of Political Economy*, 98(4).

Ahmadpoor, M. and Jones, B. (2017). The Dual Frontier: Patented Inventions and Prior Scientific Advance. *Science*, 357(6351):583–587.

Arora, A., Belenzon, S., Cioaca, L., Sheer, L., and Shvadron, D. (2023). DISCERN 2.0: Extending and Enhancing the DISCERN Dataset. *Working paper*.

Arora, A., Belenzon, S., Kosenko, K., Suh, J., and Yafeh, Y. (2024). The Rise of Scientific Research in Corporate America. *Organization Science*, 36(4):1466–1488.

Arora, A., Belenzon, S., and Sheer, L. (2021). Knowledge Spillovers and Corporate Investment in Scientific Research. *American Economic Review*, 111(3):871–898.

Bikard, M. and Marx, M. (2020). Bridging Academia and Industry: How Geographic Hubs Connect University Science and Corporate Technology. *Management Science*, 66(8):3295–3798.

Bush, V. (1945). As We May Think. *The Atlantic*.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. 134(3):1405–1454.

Clayton, P., Feldman, M., and Lowe, N. (2018). Behind the scenes: Intermediary organizations that facilitate science commercialization through entrepreneurship. *Academy of Management Perspectives*, 32(1):104–124.

Cohen, W., Nelson, R., and Walsh, J. (2002). Links and Impacts: The Influence of Public Research on Industrial R&D. *Management Science*, 48(1):1–23.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35(1):128–152.

Esposito, C. (2026). Global Science Sustains U.S. Innovation. *Working Paper*.

Fleming, L., Greene, H., Li, G., Marx, M., and Yao, D. (2019). Government-Funded Research Increasingly Fuels Innovation. *Science*, 364(6446):1139–1141.

Fleming, L. and Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, 25(8-9):909–928.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Jaffe, A. (1989). Real Effects of Academic Research. *American Economic Review*, 79(5):957–970.

Koffi, M. and Marx, M. (2026). Cassatts in the Attic: Is There a Gender Gap in the Commercialization of Science. *American Economic Journal: Applied Economics*, Forthcoming.

Marx, M. and Fuegi, A. (2020). Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles. *Strategic Management Journal*, 41(9):1572–1594.

Marx, M. and Hsu, D. (2022). Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from "Twin" Discoveries. *Management Science*, 68(2):1330–1352.

Masclans, R., Hasan, S., and Cohen, W. (2025). Measuring the commercial potential of science. *Strategic Management Journal*, 46(9):2199–2236.

Mowery, D., Nelson, R., Sampat, B., and Ziedonis, A. (2001). The growth of patenting and licensing by U.S. universities: an assessment of the effects of the Bayh–Dole act of 1980. *Research Policy*, 30(1):99–119.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy*, 19(2):165–164.

Samila, S. and Sorenson, O. (2010). Venture Capital as a Catalyst to Commercialization. *Research Policy*, 39(10):1348–1360.

Scharfmann, E., Marx, M., and Fleming, L. (2025). Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science*, 390(6776):891–893.

Siegel, D., Waldman, D., Atwater, L., and Link, A. (2003). Commercial knowledge transfers from universities to firms: improving the effectiveness of university–industry collaboration. *Journal of High Technology Management Review*, 14(1):111–133.

Stern, S. (2004). Do Scientists Pay to Be Scientists? *Management Science*, 50(6):709–853.

Stuart, T. and Ding, W. (2006). When Do Scientists Become Entrepreneurs? The Social Structural Antecedents of Commercial Activity in the Academic Life Sciences. *American Journal of Sociology*, 112(1):97–144.

Valero, A. and Van Reenen, J. (2019). The Economic Impact of Universities: Evidence from Across the Globe. *Economics of Education Review*, 68:53–67.

Van de Ven, A., Polley, D., Garud, R., and Venkataraman, S. (2008). *The Innovation Journey*. Oxford University Press.

Werth, B. (1994). *The Billion-Dollar Molecule*. Touchstone, New York.

A. Representation of Scientific Fields in the Dataset

Table A1: Trajectory Counts by Field

		Hubs		Corporations	
Field	**Field name**	**Treated**	**Control**	**Treated**	**Control**
11	Medicine	999	999	241	241
12	Biochemistry, Genetics & Molecular Biology	119	119	15	15
13	Biology	4,517	4,517	1,020	1,020
14	Pharmacology, Toxicology & Pharmaceutics	222	222	58	58
15	Neuroscience	110	110	37	37
16	Earth Sciences	973	973	317	317
17	Physics	1,819	1,819	1,087	1,087
18	Immunology & Microbiology	249	249	72	72
19	Agricultural & Biological Sciences	1,185	1,185	136	136
20	Health Professions	539	539	83	83
21	Dentistry	97	97	7	7
22	Materials Science	3,247	3,247	1,282	1,282
23	Chemistry	1,667	1,667	226	226
24	Psychology	765	765	201	201
25	Chemical Engineering	1,331	1,331	373	373
26	Economics, Econometrics & Finance	700	700	157	157
27	Engineering	12,126	12,126	2,843	2,843
28	Environmental Science	1,532	1,532	226	226
29	Energy	141	141	29	29
30	Decision Sciences	124	124	58	58
31	Computer Science	2,650	2,650	463	463
32	Mathematics	1,027	1,027	131	131
33	Social Sciences	1,158	1,158	110	110
34	Arts & Humanities	38	38	12	12
35	Business, Management & Accounting	114	114	28	28
36	Veterinary Science	807	807	108	108
	Total	38,256	38,256	9,320	9,320

B. Change in CP using Balanced Sample

As a robustness check, we re-estimate our main specification on the subsample of *balanced* trajectories. We define a trajectory as balanced if it is observed at every step from $s = -2$ to $s = 2$ without exiting the panel through patenting. Recall that when a trajectory reaches a patent, its sequence ends and it leaves the risk set; in the full sample, this attrition begins at $s = 1$ and accumulates by $s = 2$, so the set of trajectories contributing to the later event-time coefficients differs from the set contributing to the earlier ones. Restricting to balanced trajectories holds the composition of the sample fixed across all five steps, so that the estimated step profile reflects within-trajectory changes rather than shifts in which trajectories are observed.

Table A2 reports the results. The treatment effects are qualitatively unchanged and slightly larger than in the full sample. Transitioning to a hub raises commercialization potential by 0.049 at $s = 0$ (versus 0.043 in the full sample), and transitioning to a corporation by 0.099 (versus 0.088), with both effects persisting through $s = 1$ and $s = 2$. The pre-transition coefficients, meanwhile, move toward zero: neither the hub estimate (-0.000007, $t = 0.1$) nor the corporate estimate (-0.00003, $t = 0.2$) is statistically distinguishable from zero.

The way the balanced sample is constructed makes these patterns informative in two directions. Because balancing drops trajectories that patent early, it removes from the later-step coefficients the effects of the compositional change; the fact that the post-transition effects survive this restriction indicates that the estimated acceleration is not an artifact of differential exit from the risk set. At the same time, the balanced sample is not representative, as it retains trajectories that take longer to reach a patent. These trajectories are disproportionately drawn from more basic, upstream science. Thus, the balanced estimates should be read as applying to that slower-moving subset rather than to the full population of commercializing trajectories.

Table A2: Changes in Commercialization Potential, Balanced Trajectories

	(1) Transitions to Hubs	**(2) Transitions to Corporations**
$\beta^t_{s=-2}$	-0.0000067	-0.0000328
	(0.0000560)	(0.000135)
$\beta^t_{s=0}$	0.0491^{***}	0.0989^{***}
	(0.00287)	(0.00700)
$\beta^t_{s=1}$	0.0244^{***}	0.0762^{***}
	(0.00305)	(0.00718)
$\beta^t_{s=2}$	0.0114^{***}	0.0558^{***}
	(0.00343)	(0.00793)
Trajectory FE	Yes	Yes
Pair $\times$ step FE	Yes	Yes
Observations	130,140	21,920
Trajectories	26,028	4,384

Note: Estimates restrict the sample to balanced trajectories, those observed at every step from $s = -2$ to $s = 2$ without exiting through patenting. Coefficients β^t_s give the differential change in commercialization potential for transitioning trajectories relative to matched placebos at each step, relative to the reference step $s = -1$. All models include trajectory and matched-pair $\times$ event-time fixed effects. Standard errors, clustered at the matched pair, in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

C. The `commpipe` Package

All results in this paper are produced by `commpipe`, an open-source R package that implements the framework for any user-defined treatment. The package contains no treatment-definition logic: the user supplies a vector of treated OpenAlex paper identifiers, and the package handles everything downstream. The hub and corporate analyses above are two applications of the same code, differing only in the treated list and the set of papers excluded from the placebo pool. A single script reproduces every table and figure in this paper.

Data inputs. The package combines three files, kept separate for licensing reasons. The first, CITEGRAPH, is the shortest-path citation network connecting papers to patents (about 9.8 million papers), built from OpenAlex and the Reliance on Science data (Marx and Fuegi 2020). The second, PAPERS, holds paper attributes (publication year, field, and DOI). We will release both under a permissive license. The third is the CP measure of Masclans et al. (2025), distributed under CC BY-NC 4.0, which users download directly from its authors' repository; the package never redistributes it. At load time, the package joins CP onto papers by normalized DOI and reports the share of network papers that receive a CP value, so that silent join failures are visible.

Workflow. Each stage is a separate, inspectable function. The user loads the data (`cp_load`), resolves the treated list against the network and reports which papers are analyzable (`cp_resolve_treated`), keeps only clean entries whose immediate antecedent is untreated (`cp_filter_entries`), builds treated trajectories (`cp_trajectories`), constructs the placebo pool with optional user-supplied exclusions (`cp_control_pool`), builds placebo trajectories through the identical code path, matches on field, year, CP decile, and CP change with a caliper (`cp_match`), reshapes to a long panel (`cp_reshape_long`), and estimates the CP event study and the patenting hazard (`cp_estimate_cp`, `cp_estimate_hazard`). Helper functions produce the balance, observation, and CP-mean tables, the event-study and cumulative-patenting plots, and the exposure diagnostic shown in Figure 3, which reports how often placebo trajectories later pass through the treatment environment.

Validation. The package ships with a synthetic data generator that mimics the three inputs,

and its test suite runs the full pipeline end to end on these data. Two design checks guard against spurious results. Treated and placebo trajectories are built through a single code path with identical tie-breaking rules, because asymmetric construction can generate apparent effects under random treatment assignment. Users can also run a placebo test by supplying randomly drawn papers as the treated set, for which the estimated effects should be near zero.

Dissemination. We will release the package on GitHub with documentation and a worked example, deposit the CITEGRAPH and PAPERS files on Zenodo, and provide instructions for obtaining the CP data.