Global Science Sustains U.S. Innovation

Christopher Esposito
UCLA
christopher.esposito@anderson.ucla.edu

July 14, 2025

Abstract

Like physical products, new technologies are developed using globally sourced inputs. Yet while the supply chains behind physical products are well understood, we know far less about the international "supply chain" of scientific knowledge that powers U.S. innovation – and how vulnerable it is to disruption. Here, I assess the resilience of the U.S. knowledge supply chain by tracing multi-generational citation paths that connect NSF-funded research to downstream patents along their shortest citation paths. By analyzing how these paths adapt to simulated barriers impeding the flow of knowledge across the U.S. border, I evaluate the reliance of U.S. patenting on global science. I find that the paths linking NSF-funded research to patents rely heavily on foreign contributions: 52% of the intermediary papers that link NSF-funded research to patents are produced outside the U.S. Restricting knowledge flows across the U.S. border reduces the connectivity of these paths, extends their length, and lowers innovation productivity as measured by the U.S. patent-to-publication ratio. Most consequentially, border restrictions trap promising knowledge trajectories outside U.S. borders: I estimate that there are currently 11,742 NSF-stimulated paths being developed outside the U.S. While 7,095 of those paths are expected to return to the U.S. for patenting if cross-border knowledge flows were not restricted, virtually none would under a regime of U.S. scientific autarky.

Modern technologies often build on scientific insights. However, the journey from discovery to application is rarely confined to one country. Take CRISPR, a gene-editing technology, as one example: foundational research and genome cataloging were funded by U.S. agencies such as the NIH and NSF in the 1990s and early 2000s [1, 2, 3], and conceptual breakthroughs were subsequently made in Spain and France [4, 5, 6]. Yet, the resulting tools were ultimately patented in the U.S. [7, 8]. This example is not unusual. Because scientific research involves global participation, discoveries and their ensuing trajectories of research and development may cross international borders multiple times before they are applied in technology, mirroring in many ways the structure of the global supply chains for physical products.

Just as trade wars and export controls have exposed the fragility of global supply chains for physical goods, recent geopolitical shifts are testing the resilience of the global supply chain of scientific knowledge. Governments are increasingly asserting control over strategic technologies, tightening visa regimes, and redirecting research investments to serve national priorities [9, 10], and in some cases are taking adversarial stances toward their own national research institutions. Already, academic conferences have been relocated and their participation affected by restrictions in crossing international borders [11]. Whether enforced through law, norms, or fear, the resulting barriers to the movement of people reduces the ability for scientists to interact with physical proximity and transmit ideas [12, 13]. These developments have renewed urgency around fundamental questions: how dependent is U.S. innovation on international scientific collaboration? And what are the consequences when flows of science across the U.S. border are disrupted?

To evaluate the resilience of U.S. innovation to disruptions in its global knowledge supply chain, this study traces multi-generational citation paths connecting NSF-funded papers to downstream U.S. patents via their shortest paths. These paths, which often span multiple countries and years of research, reveal how ideas circulate through the global scientific ecosystem before contributing to domestic innovation. To anticipate how the connectivity of NSF investments to downstream patents would change under restrictions to the movement of knowledge across the U.S. border, I simulate disruptions to international knowledge exchange by probabilistically removing citations that cross the U.S. border and allowing the shortest-path network to reconfigure along its remaining citation edges. These reconfigured paths serve to illustrate the viability of alternate routes from NSF-funded research to patents.

The analysis contains three main parts. First, it describes the normal pathways by which knowledge advances from NSF-supported papers to downstream patents under the status quo, with no added impediments to cross-border knowledge flow. Next, it examines how increasing restrictions at the U.S. border would cause these paths to reconfigure and, in turn, affect the efficiency of the U.S. innovation ecosystem. Finally, it assesses the consequences of knowledge flow restrictions for outstanding paths that have not yet been linked to patents. Together, the results offer a quantitative stress test of the U.S. knowledge supply chain and provide new evidence on the interdependence of global science and national technological progress. An elaboration of the data sources and methods, as well as additional analyses, are provided in the Supplemental Materials (SM).

Figure 1A presents a schematic of the methods used to construct the citation network that links NSF-funded papers to downstream patents. Papers that acknowledge NSF support are taken as the "start points" for walks along the paper-patent citation graph to the nearest patent [14]. For each of these papers, I extract the citations they receive from other papers and (potentially) from patents. I repeat these steps until each walk reaches the nearest patent. Similar to [15], I retain only the shortest paths between papers and patents. In the example of Figure 1A, these methods result in a dataset containing two paths, one (in blue) originating at Paper 5 and one (in green) originating at Paper 7.

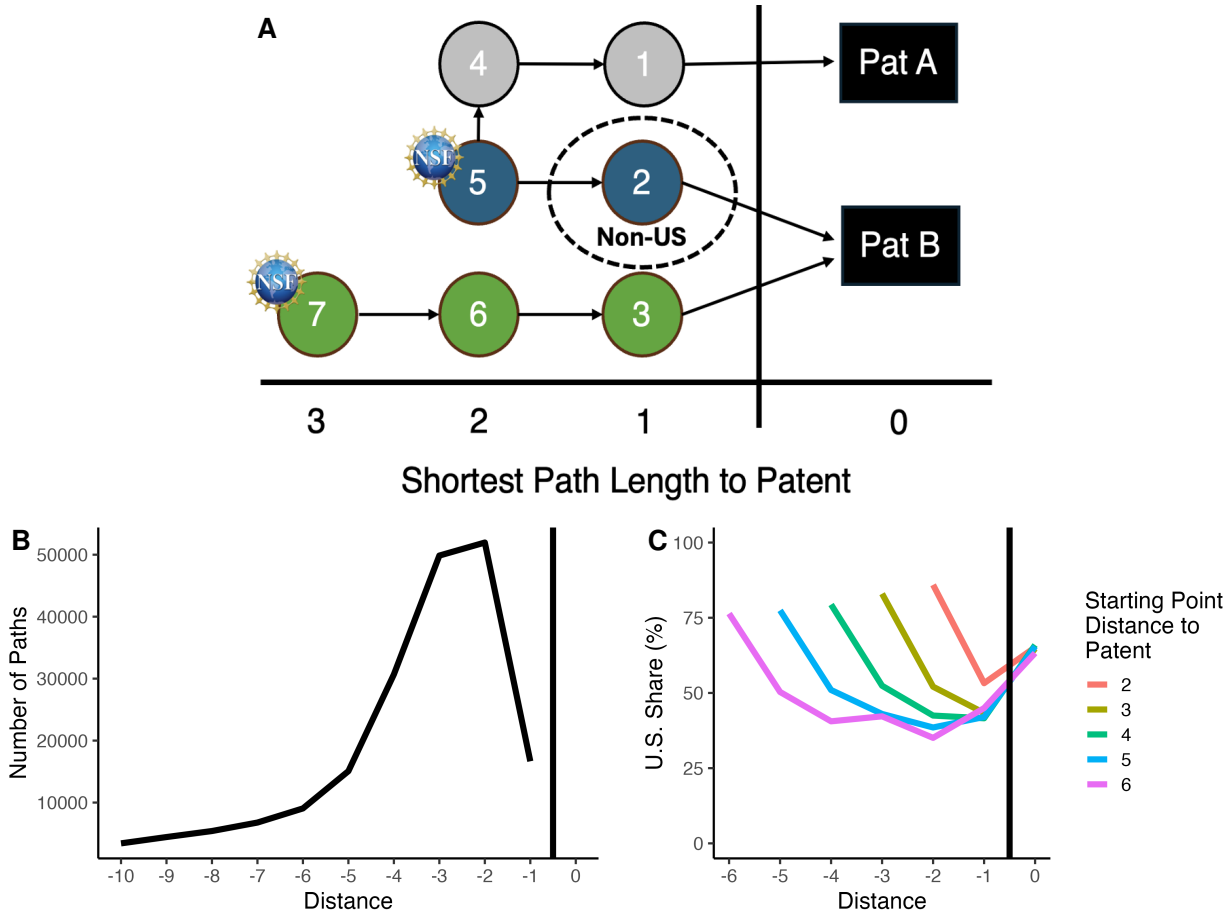


Figure 1: Citation paths connecting NSF-supported research to patents and the role of foreign contributions. (A) Shortest paths connecting NSF-funded papers to U.S. patents are extracted from the full citation network and from counterfactual networks in which citation edges crossing the U.S. border are probabilistically excluded, simulating cross-border knowledge flow restrictions [15]. **(B)** The distribution of NSF-funded papers by distance to the nearest patent shows that the modal paper is two citation steps away from a patent along the shortest path, with a long-left tail. **(C)** The share of papers produced in the U.S. by position along these paths shows that 82% of directly supported papers are U.S.-authored, as are 66% of the endpoint patents. However, only 48% of the intermediary papers connecting supported papers to their patents are U.S.-produced.

In addition to these "true" shortest paths, Figure 1A also shows how the shortest paths would reconfigure under restrictions to knowledge flow across the U.S. border. In the extreme case of U.S. scientific autarky, in which all citations that cross the border are excluded, the path beginning with Paper 5 would

re-route to pass through the grey papers (Paper 4 and Paper 1) and end at Patent A. Thus, U.S. autarky would elongate the path and cause it to end at a different patent.

My first results concern the citation distance between NSF-funded papers and their nearest downstream patents. The modal NSF-funded paper is two citation steps away from the nearest patent (Figure 1B). The distribution has a long-left tail, indicating that relatively few NSF-funded papers are at Distance values greater than 6.

To illustrate how these citation paths evolve geographically, Figure 1C shows the share of papers and patents along each path that are produced in the U.S. Paper locations are defined by the affiliation country of the last author, and patent locations are fractionally assigned based on the inventors' countries of residence (see SM for details). The vast majority of the start-points of these paths (the papers directly supported by the NSF) are produced in the U.S. However, as the paths progress toward the nearest patent, the share of U.S.-authored papers declines. For example, of the paths with stat-points at $Distance = -3$, 82% of the start-point papers are produced in the U.S., but by the time those same paths reach $Distance = -2$, only 52% of the papers are produced in the U.S.; by the time those paths advance to $Distance = -1$, only 43% at $Distance = -1$ are produced in the U.S. Remarkably, at the end of the paths, when the culminate in patents ($Distance = 0$), production shifts back to the U.S., with 65% of endpoint patents produced by U.S. inventors. This pattern suggests that while U.S.-based scientists often initiates these paths and U.S.-based inventors produce the downstream patents, international contributions advance between their start-points and end-points.

These paths can also be visualized using a flow diagram to study their production moves across countries as the paths advance from their start-points to their terminating patents. To simplify the geography of these flows, I aggregate individual countries into broader global regions based on location, participation in supranational organizations, and geopolitical relevance, as specified in the SM. I then compute the share of paths that transition between each region pair as the paths approach patents. Figure 2A presents a flow chart of paths that start at $Distance = -3$.

Figure 2A shows that while many of the shortest paths connecting NSF-supported research begin in the U.S. and terminate as patents in the U.S., they often circulate through international regions along the way. In particular, many of these paths pass through Europe, the U.K., China, East Asia, the Rest of Asia, and Canada before shifting back to the U.S. for patenting. Of the 10 non-U.S. regions, the only other region with net positive inflows of science used for patenting is Israel, a country known for its technology commercialization ecosystem.

To better understand why the paths connecting NSF-funded research to patents tend to return to the U.S. for patenting, Figure S6A in the SM provides a map of the number of captured paths by U.S. metropolitan areas. The metropolitan areas centered on San Jose, CA (Silicon Valley) and San Francisco capture the largest number of paths, followed Boston, New York, and Seattle. These metropolitan areas have advanced ecosystems for entrepreneurship and technology commercialization, which facilitate the capture of global research for technology development [16, 17, 18, 19, 20].

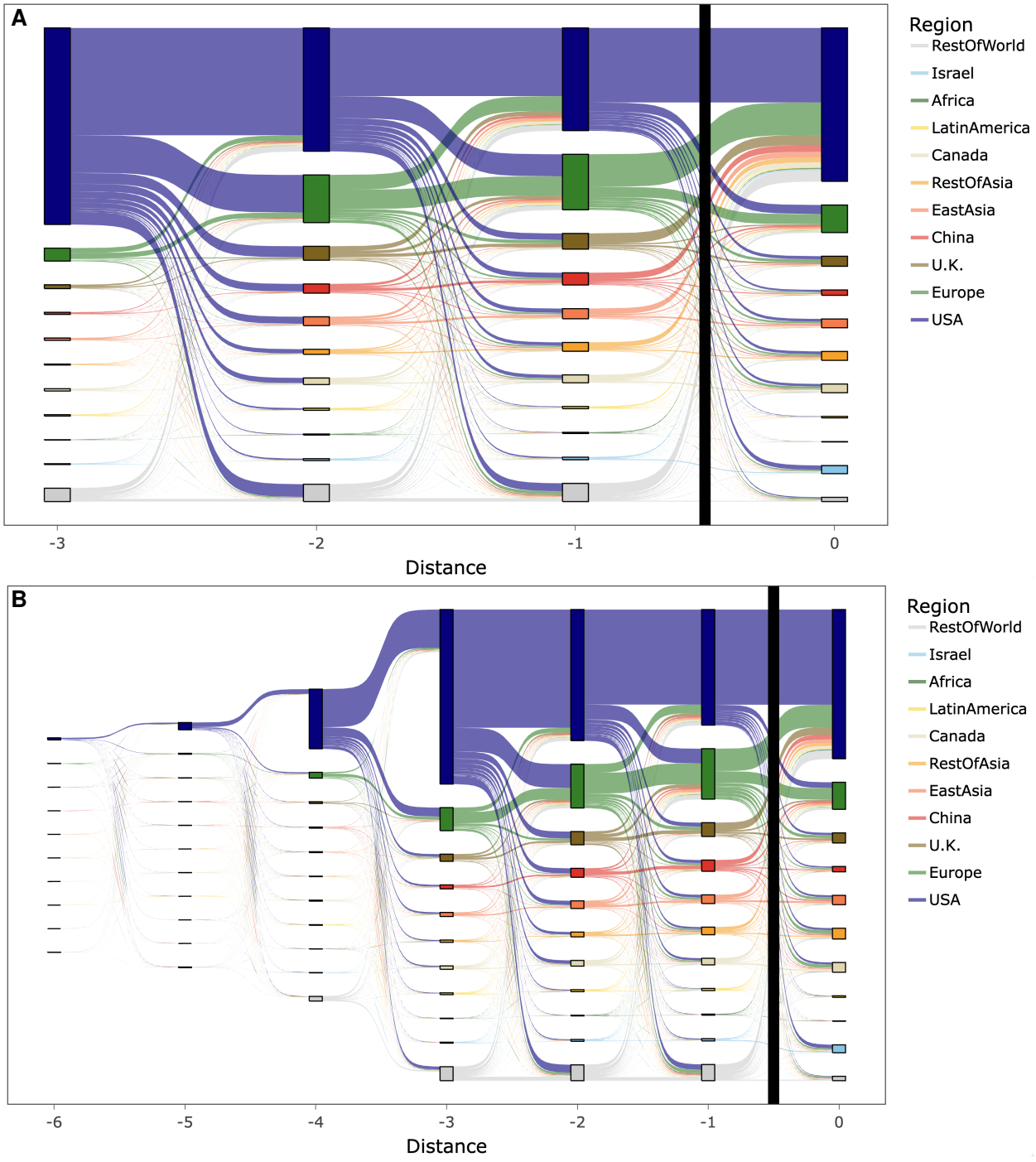


Figure 2: Geography of shortest paths from NSF-funded research (Distance = –3) to U.S. patents. (A) In the full citation network, most paths (83%) begin in the U.S., pass through international locations at $Distances = -2$ and $Distances = -1$, and return to the U.S. for patenting. **(B)** When citations crossing the U.S. border are probabilistically removed with a 50% exclusion rate, 27% of the paths start further from the patent. The share of paths crossing the U.S. border also declines.

To study how restrictions to the movement of knowledge across the U.S. border affects the global circulation of paths connecting NSF-funded research to patents, Figure 2B shows the geography of the paths that originated at $Distance = -3$ in the full citation network. However, the citation network used for Figure 2B probabilistically removes citations that cross the U.S. border with a 50% cross-border exclusion rate. The disrupted paths are allowed to reconfigure along their shortest remaining paths, if one remains. As a result of these disruptions, 27% of paths now begin more than three citation steps away from the nearest patent, as indicated by the flows positioned at Distance values of –6, –5, and –4 in Figure 2B. In addition, fewer paths leave the U.S. as they advance toward the patent boundary, and a smaller share of those that do leave ultimately return to the U.S. for patenting.

To quantify the effects of such disruptions on the connectivity of these paths, Figure 3A plots the number of captured paths that terminate at U.S. and non-U.S. patents across a full gradient of cross-border knowledge flow restrictions. In the first scenario, there are no added barriers to cross-border knowledge flow; citations crossing the U.S. border are fully preserved (0% cross-border exclusion rate). In the second through tenth scenarios, cross-border citations are probabilistically removed with progressively higher exclusion rates. In the final scenario, 100% of cross-border citations are removed, simulating a condition of full U.S. scientific autarky. These scenarios represent increasingly large barriers to knowledge flow across the U.S. border.

Figure 3A shows that under the status quo (0% cross-border exclusion), 127,936 paths terminate at U.S.-based patents, and 65,376 paths terminate at non-U.S. patents. I refer to these values as the number of *captured paths* by each region. The number of U.S. and non-U.S. captured paths declines as cross-border knowledge flow barriers increase; however, these declines are somewhat limited. In the autarky scenario (100% exclusion rate), the U.S. still captures 112,377 paths. In addition, while the non-U.S. region only captures 26,181 paths under U.S. autarky, the impact of less extreme exclusion rates on capture in the non-U.S. region are substantially smaller. For example, the 50% exclusion case results in over 60,000 paths still captured outside the U.S.

To more precisely interpret changes in path capture across scenarios, Figure 3B shows the change in the number of captured for the U.S. and non-U.S. regions, relative to the status quo (0% border exclusion rate). Notably, for intermediate levels of restriction ($\leq 60\%$ exclusion), there is a larger decline in the number of U.S.-captured paths than non-U.S. captured paths. This result arises because, under the status quo, paths captured by the U.S. cross the U.S. border on average 0.45 times, while those captured outside the U.S. cross the border on average 0.39 times (Figure S4). At the path level, the probability of disruption is an exponential function of the number of U.S. border citations. Thus, paths that cross the border more frequently face greater risk of disruption. In the extreme scenarios ($> 70\%$ exclusion rate), the decline in capture greater for the non-U.S. region, because these paths are less likely to cross the U.S. border to begin with.

In addition to these effects path capture in the U.S. and non-U.S. regions, Figure S7 in the SM disaggregates capture by individual regions, such as countries and supranational bodies. The disaggregation

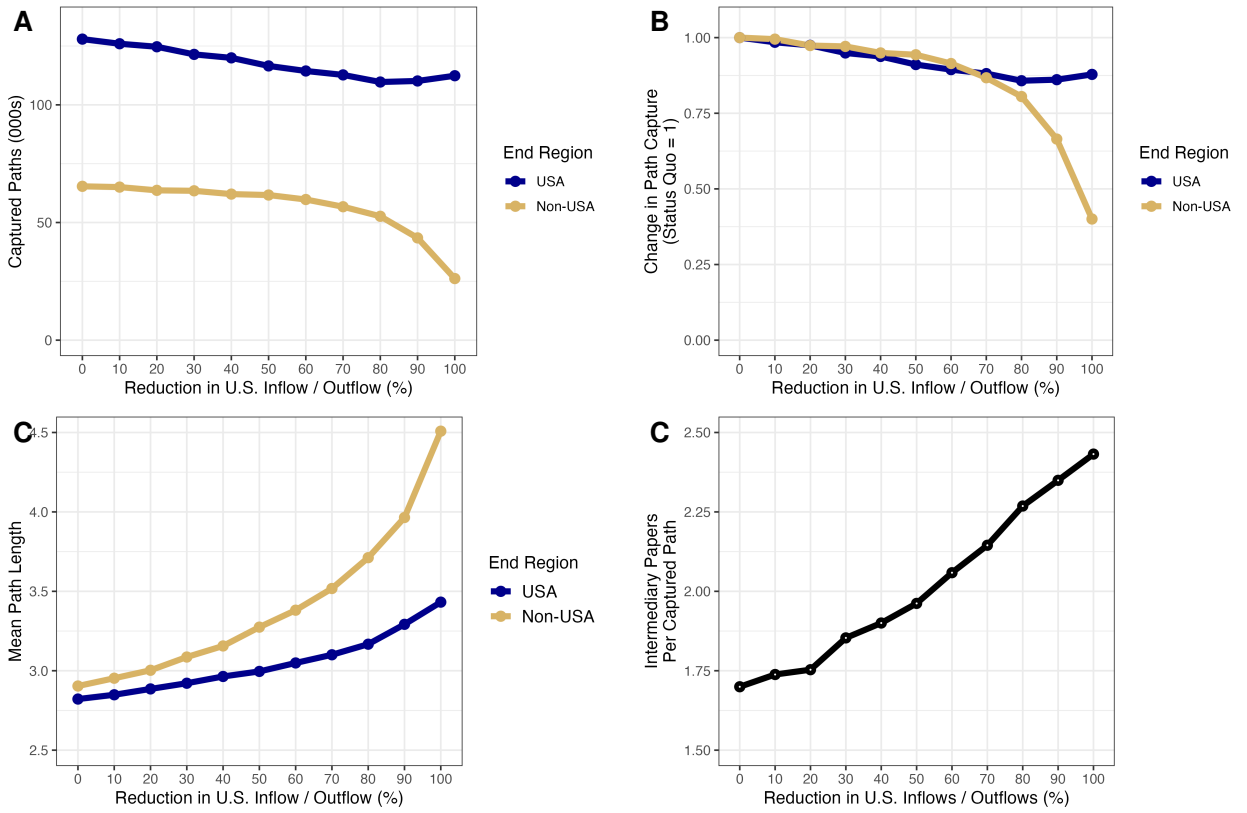


Figure 3: Effects of restrictions on knowledge flows across the U.S. border. (A) 127,936 paths terminate in U.S.-based patents under the status quo (no increase in knowledge flow barriers), and 112,377 do so under full U.S. autarky (100% reduction in U.S. inflows/outflows). **(B)** The change in the number of paths ending in U.S. and non-U.S. patents shows that the number ending in the U.S. initially decreases more rapidly than those ending abroad, because U.S.-bound paths cross the U.S. border more frequently (see text). However, under autarky the number of paths ending in non-U.S. patents declines more sharply. **(C)** The average length of citation paths linking NSF-supported articles to patents increases as knowledge flows across the U.S. border are restricted, with greater increases in path length for those ending in non-U.S. patents. **(D)** The number of intermediary papers connecting NSF-supported science to U.S. patents that are produced in the U.S. rises with increasing restrictions on knowledge flow, suggesting reduced innovation efficiency.

shows that the status quo (no additional border restrictions) results in 21,281 paths captured by Europe, 7,523 captured by the U.K., 3,763 captured by China, and 6,821 captured by Israel. The figure also shows that larger barriers to knowledge flow across the U.S. border are particularly reduce Israeli path capture, which declines by 70% under U.S. autarky.

To explore how border restrictions affect innovation productivity, Figure 3C shows the relationship with cross-border citation exclusion rates and the change in path length. As Figure 1A demonstrated, a path can increase in length if it contained an excluded cross-border citation and the shortest alternative path requires more citation steps to reach a patent. Figure 3C shows that, under the status quo of no new restrictions, mean path length is 2.82 for U.S.-captured paths and 2.90 for non-U.S.-captured paths. As cross-border barriers increase, path lengths grow in both cases – eventually reaching 3.43

for U.S.-captured paths and for non-U.S. captured 4.51 paths.

Why do paths ending outside the U.S. increase in length more than those that end in the U.S.? To interpret this finding, Figure S5 plots the increase in path length for U.S. and non-U.S.-ending paths by the Distance to the nearest patent of their starting points (their NSF-funded papers). The figure shows that the difference in the increase in path length for U.S. and non-U.S. ending paths is driven disproportionately by paths that start near the paper/patent boundary. Comparing the status quo to full U.S. autarky, paths that start at at $Distance = -1$ increase in length by 21% when ending in the U.S., but by 84% when ending outside the U.S. In contrast, paths starting at $Distance = -5$ increase by similar amounts regardless of endpoint. Therefore, the larger overall increase in path length for non-U.S.-ending paths under U.S. autarky appears to reflect greater difficulty for non-U.S. actors to translate scientific knowledge into patents when flows with the U.S. are disrupted.

In addition to lengthening citation paths, restrictions on cross-border knowledge flow can alter the innovation productivity of regions, defined as their ratio of the number of intermediary papers produced to the number of captured paths. This ratio can change through two mechanisms. First, as path lengths increase, more intermediary papers connecting the starting points to patents are required. Some of these additional intermediary papers may be produced in the focal region, leading to a higher regional intermediary papers / captured path ratio. Second, under tighter cross-border restrictions, a greater share of intermediary papers may be produced within a region as regional scientists and inventors become less able to draw on foreign contributions. Thus, the regional composition of the intermediary papers in a path can change.

Putting these two mechanisms together, Figure 3D shows the intermediary paper / captured path ratio for U.S.-captured paths. Under the status quo, the U.S. produces 1.7 intermediary papers per captured path. As cross-border restrictions grow, this ratio increases, reaching 2.43 under full U.S. autarky. This 43% increase suggests an increase in the intensity of domestic scientific effort required to sustain the same level of connectivity to downstream patents.

While restrictions to knowledge flows across the U.S. border would lengthen the length of paths connecting NSF-funded research to downstream patents and increase the input intensity of U.S. path capture, it would also have more immediate impacts on the outstanding paths that have not yet connected to patents. There are 56,341 NSF-supported papers in the data sample that have not yet been linked to a patent. The eventual outcomes of their paths, including whether, when, and where they connect to a patent, will be shaped by changes in the ability to transmit knowledge across the U.S. border.

Forecasting how such restrictions would affect the outcomes of outstanding paths requires several inputs. First, it requires an estimate of how many of these paths would ultimately link to a patent under the status quo of no new border restrictions. Second, it requires estimates of the citation distance from starting point of the path (the originating NSF-funded paper) to the nearest patent (hereafter *StartingD*). Third, it requires estimates of the current citation distance of each path to a patent, given that paths may have advanced closer to patents in the years since they originated (hereafter *CurrentD*).

Fourth, it requires information on the current region where each path is presently under development (hereafter *CurrentRegion*). Finally, for each cell in this $(\mathrm{StartingD}, \mathrm{CurrentD} = d, \mathrm{CurrentRegion} = r)$ space, forecasting the impact of border restrictions on the outcomes of outstanding paths requires a probabilistic estimate as to whether, when, and where each path will eventually link to a patent. These estimates must be made for a range of scenarios, from the status quo through full U.S. scientific autarky

To compute these estimates, I use information from the paths that have already been linked to patents, which I call realized paths. First, to estimate the number of outstanding paths that will eventually link to patents, I use information on the number of years elapsed between the start of realized paths (i.e., the publication year of the originating NSF-supported article) and the year the path is realized via a patent. I use this information to down-sample the outstanding paths based on their current path age. Second, I assign StartingD values to outstanding paths by using the joint distribution of years-to-realization and StartingD from the realized paths to probabilistically assign StartingD values based on each path's age. Third, I assign CurrentD values to outstanding paths based on the joint distribution of CurrentD within StartingD bins for realized paths and the corresponding values for the outstanding paths. Finally, I probabilistically assign CurrentRegion values to outstanding paths based on their StartingD and CurrentD values and the observed frequency of realized paths with the same values across regions. (See SM for details on methods).

After assigning each outstanding path to an $(\mathrm{StartindD}, \mathrm{CurrentD} = d, \mathrm{CurrentRegion} = r)$ cell, I use the realized paths to compute transition probabilities from each cell to each possible outcome. To explore how cross-border knowledge restrictions would affect these outcomes, I compute these probabilities using the full shortest-path network (the status quo), the U.S. autarky scenario, and the nine intermediate levels of border restrictions.

More formally, for each $(\mathrm{StartingD}, \mathrm{CurrentD} = d, \mathrm{CurrentRegion} = r)$ cell and under each knowledge flow restriction scenario, I compute the probability that a path has the following outcomes:

1. Terminates at a patent in region r with CurrentD of d
2. Terminates at a patent in different region $r' \neq r$ or at a greater distance $d' > d$
3. Stalls (does not terminate at a patent)

Figure 4A shows the number of outstanding paths captured by the U.S. and outside the U.S. by cross-border citation exclusion rates, and Figure 4B shows the change in these totals as cross-border knowledge flow restrictions increase. Under the status quo, the U.S. captures 15,167 outstanding paths and the non-U.S. region captures 7,928 paths. These counts decline as restrictions rise, ultimately reaching 9,884 U.S.-captured paths and 4,611 non-U.S.-captured paths in the autarky scenario. Thus, relative to the status quo, U.S. autarky results in 5,283 fewer outstanding paths captured by the U.S.

The economic significance of this decrease in path capture can be interpreted in terms of cost of the NSF's investments in scientific research. The NSF invests between $4 billion and $10 billion annually

in basic research, totaling roughly $150 billion since 2000. Over that period, 233,842 papers with 10 or more forward citations have acknowledged NSF funding. Assuming that NSF funding was evenly distributed across these papers ($640,000 per paper), a loss of 5,283 captured paths implies $3.4 billion in NSF-supported research that is no longer domestically captured by U.S.-based patents.

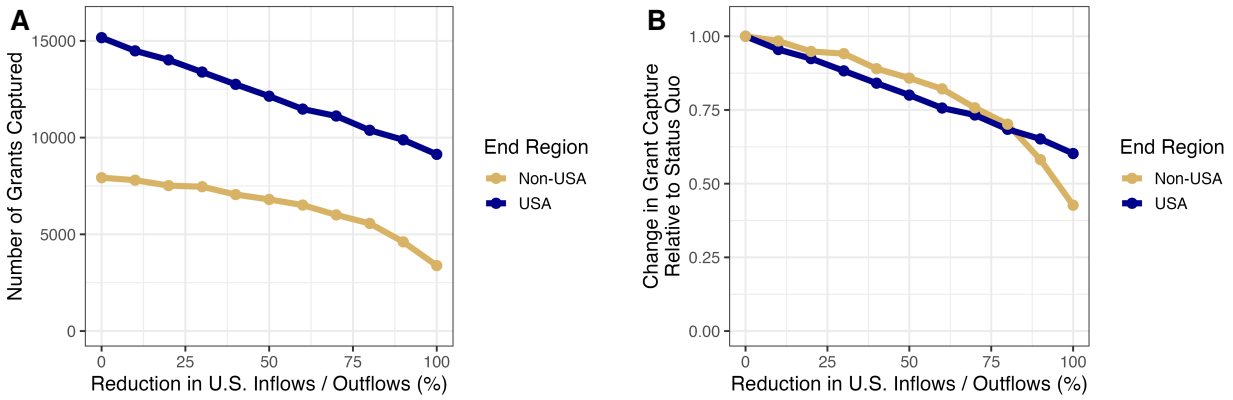


Figure 4: Capture of outstanding paths. (A) Number of outstanding paths captured in the U.S. and outside the u.S. under each level of cross-border knowledge flow restrictions. **(B)** Change in the number of captured outstanding paths shows that the number ending in both U.S. and non-U.S. patents declines as barriers to knowledge flow increase, and that the number ending in U.S. patents declines more sharply at intermediate levels of restriction.

Figure 4B shows that intermediate cross-border knowledge flow restrictions (i.e., reduction in flows $< 80\%$) lead to a larger reduction in the number of outstanding paths ending in U.S. patents than in non-U.S. patents. This result arises because many of the outstanding paths (51% of the total) are assigned to non-U.S. regions (see SM Figure S3). In other words, many outstanding paths have already "left" the U.S. knowledge system and are currently being advanced by foreign scientists. As barriers to cross-border knowledge flow increase, the share of outstanding paths that return to the U.S. for patenting declines, reflecting a loss of downstream capture for U.S.-funded research.

While this study generates an analysis of the potential impact of cross-border knowledge flow restrictions on U.S. innovation, it has some limitations. First, the analysis traces knowledge flows using the shortest paths between papers and patents [15]. This method has a key advantage in that enables efficient counterfactual modeling, as captured by shortest remaining paths in restricted networks. However, the shortest-path method also introduces an important limitation in that it does not describe a generative process of knowledge creation. For example, one might reasonably believe that all citations made by a focal paper are necessary for that paper's development. As a result, this analysis may underestimate the true impact of knowledge flow barriers on U.S. innovation.

Second, currently-available data may understate the importance of Chinese science to U.S. innovation. Chinese science has rapidly grown in scale, sophistication, and global leadership [9]. While U.S. patents are increasingly citing Chinese science directly [21], it remains to be seen whether this increase in activity is also generating indirect effects through multi-generational citation paths.

References

[1] S.F. Altschul, W. Gish, E.W. Myers, and D.J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] R.L. Tatusov, E.V. Koonin, and D.J. Lipman. A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637, 1997.

[3] Dennis Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David Lipman, James Ostell, and Eric Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36–D42, 2013.

[4] Francisco Mojica, Cesar Diez-Villasenor, Elena Soria, and Guadalupe Juez. Biological Significance of a Family of Regularly Spaced Repeats in the Genomes of Archaea, Bacteria and Mitochondria. *Molecular Microbiology*, 36(1):244–246, 2000.

[5] Ruud Jansen, Jan van Embden, Wim Gaastra, and Leo Schouls. Identification of Genes that are Assoicated with DNA Repeats in Prokaryotes. *Molecular Microbiology*, 43(6):1565–1575, 2002.

[6] Francisco Mojica, Cesar Diez-Villasenor, Jesus Garcia-Martinez, and Elena Soria. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 60:174–182, 2005.

[7] Feng Zhang. CRISPR-Cas Systems and Methods for Altering Expression of Gene Products, 2014.

[8] Jennifer Doudna, Martin Jinek, Charpentier Emmanuelle, and Krzysztof Chylinski. Methods and Compositions for RNA-Directed Target DNA Modification and for RNA-Directed Modulation of Transcription, 2019.

[9] Renli Wu, Christopher Esposito, and James Evans. China's Rising Leadership in Global Science. *arXiv: 2406.05917*, 2024.

[10] Aaron Chatterji and Fiona Murray. How Geopolitics Is Changing the Economics of Innovation. In *Entrepreneurship and Innovation Policy and the Economy*, volume 5 of *NBER Book Series*. University of Chicago Press, 2025.

[11] Miryam Naddaf. Scientific Conferences are Leaving the U.S. amid Border Fears. *Nature*, 2025.

[12] Yiling Lin, Carl Benedikt Frey, and Lingfei Wu. Remote Collaboration Fuses Fewer Breakthrough Ideas. *Nature*, 623:987–991, 2023.

[13] Benjamin Balsmeier, Lee Fleming, and Sonja Lück. Isolating Personal Knowledge Spillovers: Coinventor Deaths and Spatial Citation Differentials. *American Economic Review: Insights*, 2023.

[14] Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. SciSciNet: A Large-Scale Open Data Lake for the Science of Science Research. *Scientific Data*, 10(215), 2023.

[15] Mohammad Ahmadpoor and Benjamin Jones. The Dual Frontier: Patented Inventions and Prior Scientific Advance. *Science*, 357(6351):583–587, 2017.

[16] Sampsa Samila and Olav Sorenson. Venture Capital as a Catalyst to Commercialization. *Research Policy*, 39(10):1348–1360, 2010.

[17] Sampsa Samila and Olav Sorenson. Venture Capital, Entrepreneurship, and Economic Growth. *Review of Economics and Statistics*, 93(1):338–349, 2011.

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

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

[20] Josh Lerner, Henry Manley, Carolyn Sein, and Heidi Williams. The Wangering Scholars: Understanding the Heterogeneity of University Commercialization. *NBER Working Paper Series*, (32069), 2024.

[21] Junhan Wang, Xibao Li, and Yangbo Wang. The Contribution of Chinese Science to US Technological Advancement: Evidence from Patent Citation of Academic Papers. *SSRN Working Paper #4802379*, 2024.

[22] Matt Marx and Aaron Fuegi. Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles. *2020*, 41(9):1572–1594, 2020.

Supplemental Materials

1. Materials and Methods

1A Data Sources

My analysis combines patent records with data on scientific publications. The patent data comes from three sources. First, from the USPTO (via PatentsView), I source patent numbers and inventor location information. Second, from the Reliance on Science dataset [22], I source citation records on patents made to scientific articles. Third, from PatStat, I source patent family information. I use the patent family records to identify patents that are applied with the three major patent offices (the USPTO, JPO, and EPO), and I restrict my study to these "triadic" patents in order to reduce home-market biases. Furthermore, to reduce potential volatility in the dataset due to patents that are weakly connected to science, I restrict the patent sample to patents that cite 5 or more scientific articles. This results in a sample of 301,237 patents.

For data on scientific publications, I retrieve records from two sources. First, I retrieve paper numbers and citation records from OpenAlex. Second, I source author affiliation information (including location information) from SciSciNet [14], a data lake that cleans and combines publication information from a variety of open-access sources, particularly OpenAlex. I retrieved affiliation information from SciSciNet rather than directly through OpenAlex because the authors of SciSciNet dataset have taken measures to improve the quality of the underlying data (14). I retain only the last authors on scientific papers, as a proxy for the principal investigators. Thus, I assign papers to countries based on the affiliation location of the last authors. I remove all papers for which SciSciNet does not provide an affiliation country for the last author.

Finally, from SciSciNet I source acknowledgments of NSF funding support, at the paper level [14]. Articles that acknowledge NSF support serve as the "start points" when identifying the citation paths that trace from NSF-funded research to downstream patents. To ensure that my dataset does not include non-substantive papers acknowledging NSF support, I restrict the set of NSF-supported articles to those that receive 10 or more forward citations.

1B Construction of Citation Paths

I construct shortest-path networks that link NSF-supported papers to their nearest patents. To do so, I begin with USPTO records of triadic patents, front-page citations from patents to scientific articles [22], and citations from scientific articles to other articles (OpenAlex). I call this complete network of all front-page patent-to-paper and all paper-to-paper citations the full citation network.

Next, I identify the scientific papers directly cited by the triadic patents. These direct citations are reflected by the patents (black squares) in Figure 1A of the main text. In the figure, the patents receive

inward-pointing edges from papers, Papers 1, 2, and 3. Because these papers are directly cited by patents, they are positioned at a value of 1 on the x axis, indicating that the shortest path length from these papers to the nearest patent is 1.

After identifying the papers that are directly cited by patents, I identify the papers that are indirectly cited by patents. To do so, I begin with the papers that are directly cited and record the papers that they cite: Papers 4, 5, and 6. These papers are located at a value of 2 on the x axis, because the shortest length path from them to the nearest patent is 2.

Next, I identify the papers at shortest length distance 3 from the nearest patent. To do so, I extract the citations made by papers at $Distance = 2$. In Figure 1, there are two papers cited by papers at $Distance = 2$: Papers 5 and 7. However, Paper 5 was also cited by Paper 2. Since Paper 5 is cited by both Paper 4 and Paper 2, and Paper 2 lies on a shorter path to the patent, the citation from Paper 5 to Paper 4 does not fall on the shortest path. Therefore, I omit the Paper 5 -¿ Paper 4 citation when computing the shortest paths.

I repeat the above steps until I arrive at scientific papers that are a distance of 10 from the nearest patent. There are a total of 16.4 million resulting scientific papers in the citation graph, linked to 165,016 patents.

After constructing the shortest-path network, I extract the paths of the network that begin with papers that acknowledge NSF support. There are 449,947 total articles that acknowledge NSF support in the SciSciNet dataset, of which 236,400 receive 10 or more forward citations. I conduct walks along the shortest-path network from each of these 236,400 articles, moving forward across the citation network until I reach a patent. When more than one equally-short walk is possible between an NSF-supported article and a patent, I randomly select one of them. In the status quo scenario, in which there are no new barriers to knowledge flow across the U.S. border, 180,059 of the 236,400 included originally-linked NSF-funded articles (76%) are linked to downstream patents.

1C Methods for simulating knowledge flow barriers at the U.S. border

To simulate barriers to knowledge flows crossing the U.S. border, I first identify citations in the full citation network where either the cited paper is outside the U.S. and the citing paper is in the U.S., or the cited paper is in the U.S. and the citing paper is outside the U.S. Next, I remove each of these cross-border citations with a set probability, producing multiple restricted networks. The 0% exclusion network removes no citations that cross the U.S. border; I call this network the status quo. The restricted networks remove citations crossing the U.S. border with 10%, 20%, ..., up to 100% probability. After constructing these networks, I identify the shortest paths connecting NSF-supported articles to patents in each of them by repeating the steps outlined in Methods Section 2.

1D Methods for down-sampling outstanding paths

To forecast where the outstanding paths (NSF-funded articles that have not yet been linked to a patent) will be captured, I first down-sample the full set of these outstanding start-points. I conduct this down-sampling because not all of the outstanding paths will eventually become linked to a patent. To down-sample the outstanding paths, I use a hazards model to probabilistically remove outstanding paths based on their age and the distribution of the age of realized paths as of the year that they first become linked to patents.

More specifically, I first identify all realized paths (those that have already linked to a patent), and record the number of years between the publication of their start-paper and their downstream patent (i.e. years-to-realization). Next, I Identify all outstanding paths and calculate their age as of 2021 (the last year covered by my patent data). Finally, I down-sample outstanding paths based on their age, using the empirical distribution of years-to-realization from the realized paths. For example, 99% of realized paths were linked to patents by age 17, so outstanding paths older than 17 years are unlikely to connect to a patent in the future. These outstanding are probabilistically dropped from the sample based on the observed hazards function from the realized paths. Figure S1 shows how the outstanding paths are down-sampled based on the keep probability calculated from the realized paths.

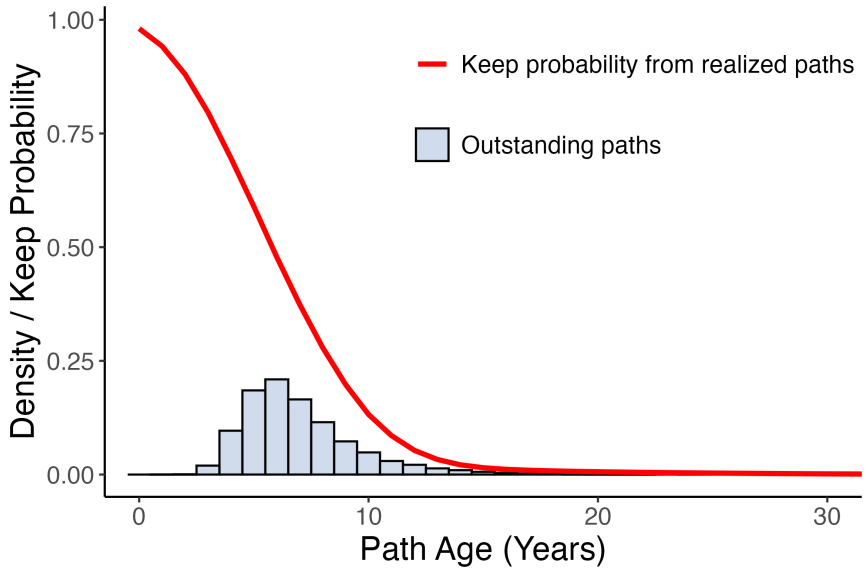


Figure S1: Down-sampling outstanding paths based on the keep probability. The keep probability is calculated using the inverse hazards function of realization with a patent, calculated using the realized paths. The resulting keep probability is used to down-sample the outstanding paths.

The down-sampling process reduces the number of outstanding paths from 56,341 to 23,268. The median path age of the surviving outstanding paths (bench-marked to 2021, the final year of the patent-to-paper citation dataset) is 6 years.

1E Methods for assigning outstanding paths to (NSF-D, Paper-D, Region) cells

After down-sampling the outstanding paths, I predict three features for each outstanding path: the start paper's distance to the nearest patent (*StartingD*), the current distance of the path (as of 2021) to the nearest patent (*CurrentD*), and the current region where the path is being advanced (*CurrentRegion*). I make these predictions based on the path age of outstanding paths, and the age distribution of realized paths across (StartingD, CurrentD, CurrentRegion) cells.

To make these assignments, I first simulate the age-at-realization of outstanding paths. I simulate these values because they are not directly observable for outstanding paths. To simulate their age at realization, I draw residual durations from the distribution of realized paths' ages-at-realization minus the distribution of the current path age of the outstanding paths. I add these residuals to outstanding paths' current ages to estimate when outstanding paths are likely be realized. This sampling coerces the distribution of age-at-realization for outstanding paths to resemble the same distribution for realized paths.

Next, using the joint distribution of realized paths by StartingD and age-at-realization, I assign outstanding paths proportionally to StartingD values conditional on their simulated age-at-realization. Thus, I combine the observed age-at-realization and StartingD relationship from realized paths with the simulated age-at-realization information on outstanding paths, to assign StartingD values to outstanding paths. In general, outstanding paths with higher simulated age-at-realization values are assigned higher StartingD values (further from the nearest patent), because more years are expected to elapse between their starting point and their realization as patents.

After assigning outstanding paths StartingD values, I assign these paths CurrentD values. To do so, I first condition on outstanding and realized paths' StartingD values and their path age. Within these groups, I assign the outstanding paths to CurrentD cells based on the empirical joint distribution from realized paths. Generally, an active grant that is "old" for its StartingD value will be assigned a low CurrentD value, because older paths tend to be closer to the paper-patent boundary. For example, an outstanding path with a simulated-years-to-realization value of 15 and a path age of 14 might be assigned to StartingD = 4 (a relatively high value) and CurrentD = 1 (the lowest possible value), because 13 years have elapsed since the path started, and because the path is expected to be realized in one year.

Figure S2 shows the density of outstanding paths that are assigned to each StartingD and CurrentD value. Overall, 4.7% of paths are assigned to StartingD = -1, 29% are assigned to $StartingD = -2$, 33% are assigned to StartingD = -3, 20% are assigned to $StartingD = -4$, and a small share are assigned to higher StartingD values. For CurrentD, the majority (72%) are assigned to $CurrentD = 1$.

Finally, I assign outstanding paths CurrentRegion values by conditioning on StartingD, CurrentD, and path age, and by assigning each outstanding path to a region based on the corresponding joint distribution from the realized paths. In general, regions with many papers in realized paths with a given

set of StartingD, Current value values are more likely to be assigned to outstanding paths at the same (StartingD, CurrentD) location.

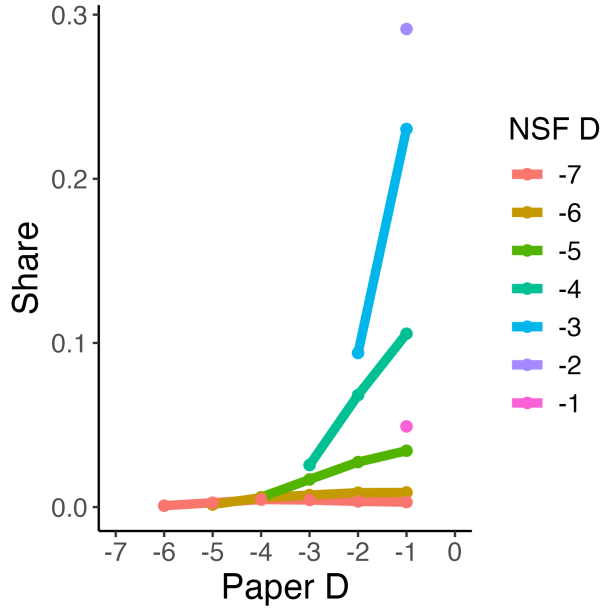


Figure S2: Assignment of outstanding paths to StartingD and CurrentD values

Figure S3 shows the assignment of outstanding paths to regions. Assigned regions are shown in lighter colors, while the region of the start-point of the paths are shown in darker colors. Both values are shown in the figure to demonstrate how the regional assignment process may assign outstanding paths to different regions from where those paths began. Moreover, most paths begin in the U.S. (see Figure 1C), when CurrentD = StartingD. However, because most of the outstanding paths are projected to no longer be at their start points (CurrentD $<$ StartingD), they are frequently assigned to regions other than their starting one.

1F Methods for computing transition probabilities

After assigning outstanding paths to (StartingD, CurrentD, CurrentRegion) cells, I project the outcomes of those paths. In particular, I estimate whether and where they are likely to end up as patents along their shortest paths, and how long those path lengths are likely to be. I estimate these outcomes in the absence of barriers to international knowledge flows (the status quo), and in the presence of such barriers.

To compute the transition probabilities in the absence of international barriers, for each $(\text{StartingD}, \text{PaperD} = d, \text{CurrentRegion} = r)$ cell, I use the outcomes of the realized paths to compute the probability that a paper terminates in a patent in region $r \in R$ at distance $d \in D$.

Next, to compute transition probabilities using the restricted networks (those with 10%, 20%, . . ., 100%

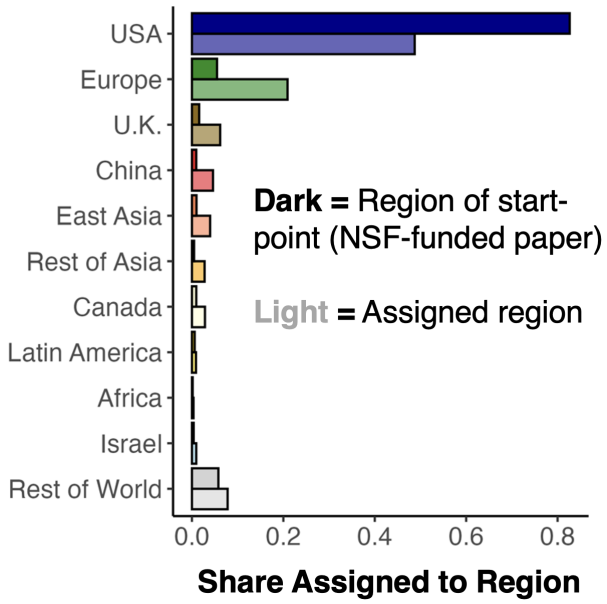


Figure S3: Assignment of outstanding paths to regions

reductions in knowledge flows across the U.S. border), for each $(\text{StartingD}, \text{PaperD} = d, \text{CurrentRegion} = r)$ cell, I use the outcomes of the papers in the realized paths to compute the probability that a paper has the following outcomes:

1. Terminates at a patent in region r with CurrentD of d
2. Terminates at a patent in different region $r' \neq r$ or at a greater distance $d' > d$
3. Stalls (does not terminate at a patent)

1G Methods for simulating outcomes of outstanding paths

After computing transition probabilities using papers in realized paths, I simulate the outcomes for outstanding paths by randomly drawing from the transition probabilities of their assigned $(\text{StartingD}, \text{PaperD} = d, \text{CurrentRegion} = r)$ cell. I perform this step using transition probabilities from both the unrestricted network and the restricted networks, generating comparable outcomes for the full gradient of possible restrictions in knowledge flows across the U.S. border.

1H Assignment of countries to regions

Countries are aggregated to regions based on their location, embeddedness in supranational bodies, and geopolitical significance. The regional definitions are as follows: **Europe** = the 27 member states of the European Union, Switzerland, and Norway. **East Asia** = South Korea, Japan, Taiwan.

Rest of Asia = Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei, Cambodia, Georgia, Indonesia, India, Iran, Iraq, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Laos, Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, North Korea, Oman, Pakistan, Palestine, Philippines, Qatar, Saudi Arabia, Singapore, Sri Lanka, Syria, Tajikistan, Thailand, Timor-Leste, Turkmenistan, United Arab Emirates, Uzbekistan, Vietnam, Yemen. **Latin America** = Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Uruguay, Venezuela, Bahamas, Barbados. **China** = Mainland China, Hong Kong, Macau. **Africa** = Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Republic of the Congo, Côte d'Ivoire, Democratic Republic of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, The Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, São Tomé and Príncipe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan, Tanzania, Togo, Tunisia, Uganda, Zambia, Zimbabwe. The **United States**, **Canada**, **U.K.**, and **Israel** are defined as independent regions. The **Rest of World** region includes all countries not mentioned in the above lists.

2. Supplemental Text

2A Overview

The supplemental text contains the following sections:

- Border crossings by end region
- Change in mean path length of realized paths by starting Distance
- Geography of endpoints of U.S.-captured paths by metropolitan area
- Analysis of regionally-disaggregated outcomes

2B Border Crossings by End Region

In Figure S4, I show the mean number times paths cross the U.S. border between their start point and end patent, broken out by their end region (U.S. or non-U.S.). Paths that end in the U.S. cross the border an average of 0.45 times, while those that end at non-U.S. patents cross the border an average of 0.39 times.

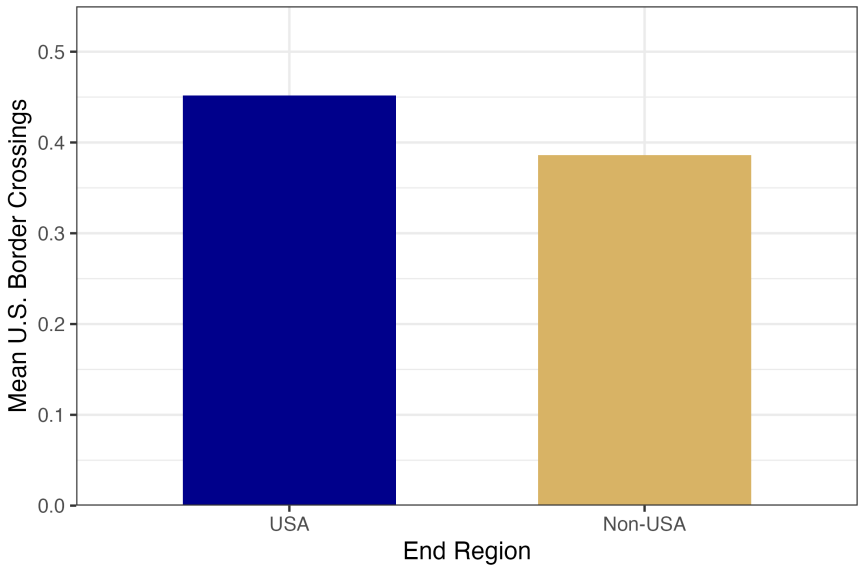


Figure S4: Mean Number of Crossings of the U.S. Border for Paths with U.S. and Non-U.S. End Regions

2C Change in Mean Path Length for Realized Paths by Starting Distance (StartingD)

The main text shows that mean path length increases for paths that end in the U.S. and outside the U.S. when barriers to crossing the U.S. border increase. In Figure S5, I decompose these mean effects by the starting Distance of the paths. This decomposition allows one to determine whether the larger increase in mean path length for non-U.S. endpoints is driven by paths that begin close to the patent boundary, or further from the boundary. Changes in the mean path length for the various knowledge flow restriction scenarios are shown for paths ending in the U.S. in Figure S5A, and for paths ending outside the U.S. in Figure S5B.

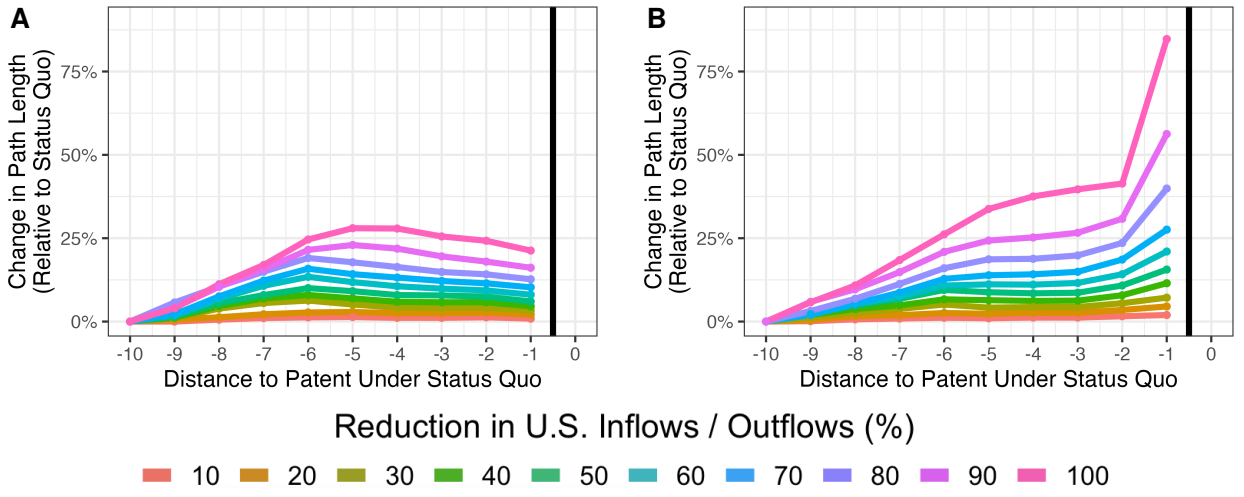


Figure S5: Change in Mean Length of Paths Connecting NSF-Supported Articles to Patents by Starting Distance. (A) Paths captured by the U.S. **(B)** Paths captured outside the U.S.

Figure S5A shows that U.S.-captured paths increase in length by about 25% between the status quo scenario and the U.S. autarky scenario for all StartingD values between -6 and -1. In contrast, Figure S5B shows that paths captured outside the U.S. increase by progressively larger amounts for paths that start closer to the patent boundary, when contrasting between the status quo and U.S. autarky cases. In particular, paths with StartingD = -1 increase in length by over 80% if they are captured outside the U.S. The much larger increase for paths that start near the patent boundary and are captured outside the U.S. suggests that paths captured outside the U.S. struggle to form new connections to patents when knowledge flows across the U.S. border are impeded.

2D Geography of Path Endpoints in the U.S. by Metropolitan Area

Figure S6 presents a map of U.S. metropolitan areas by their number of captured paths. There are very strong concentrations in the metropolitan areas centered on San Francisco and San Jose, and weaker concentrations in the metro areas centered on New York City, Boston, Seattle, Los Angeles, and San Diego. All other metro areas capture comparatively few paths. Moreover, the geography of path capture is similar to the geography of venture capital, suggesting that the United States is able to translate globally-sourced science into patents because it has agglomerations that efficiently match scientific knowledge to financial resources [16, 17].

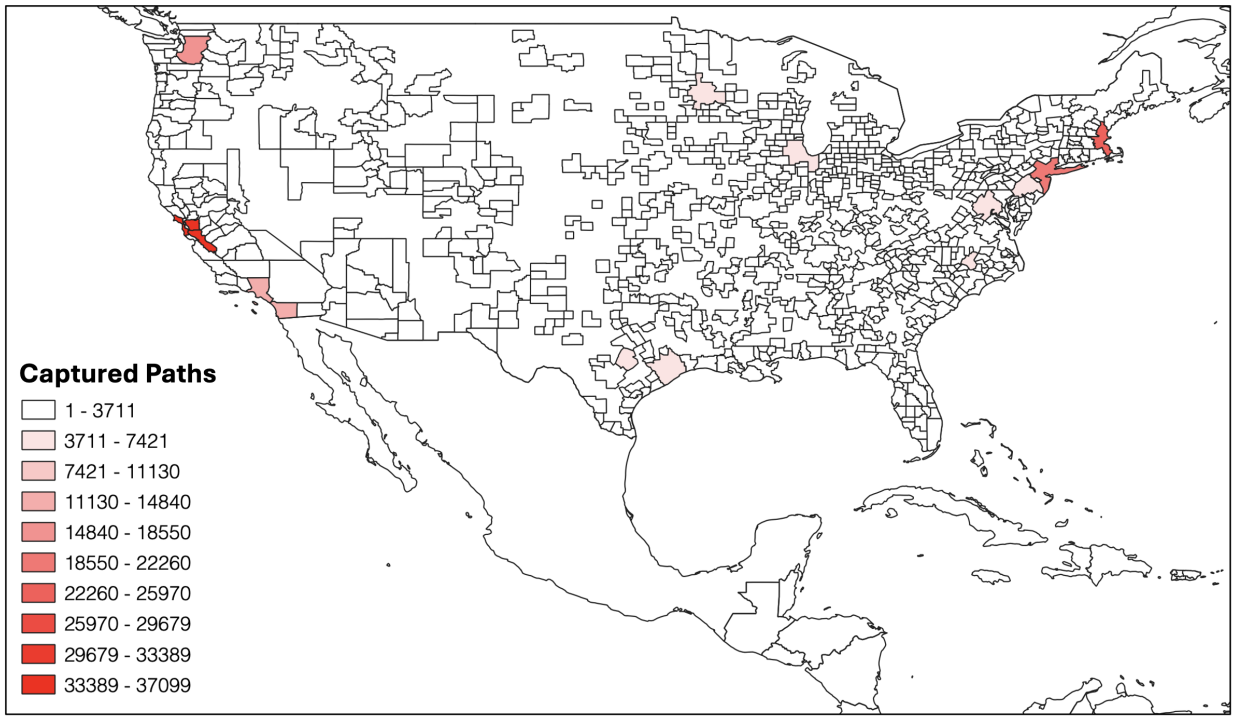


Figure S6: Number of Paths Captured by U.S. Metropolitan Area

2E Analysis of Regionally-Disaggregated Outcomes

In this section, I replicate the main analyses while disaggregating non-U.S. captured paths into their respective regions.

Figure S7 shows the change in the number of realized paths captured by each region, by knowledge flow barrier levels. Notably, knowledge barriers at the U.S. border reduce the number of paths captured by Israel, but increase the number captured in the Rest of Asia and Canada (except in the cases with very high knowledge flow barriers). The number of paths captured by Africa increases in percentage terms, but the raw number captured in Africa is very small, so this result is not meaningful.

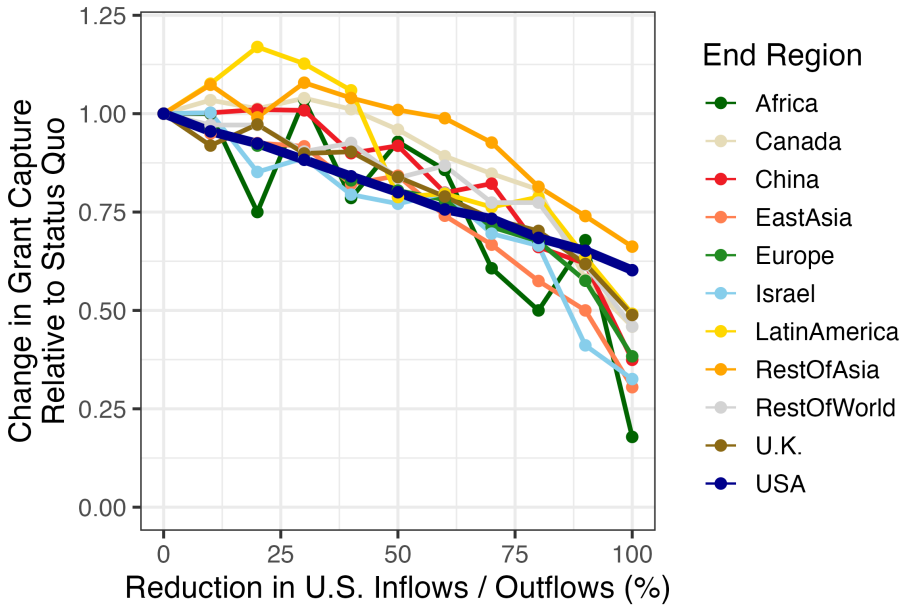


Figure S7: Change in Number of Realized Paths Captured, by All Regions

Figure S8 shows the mean length of paths by region of capture. Mean path length under the status quo is notably higher in Israel and the Rest of Asia. Mean path length increases for all regions as knowledge flow barriers at the U.S. border increase; however, paths captured by the U.S. increase in length less than do those that end in the other regions.

Figure S9 shows the mean number of crossings of the U.S. border by capturing region. Paths captured by Israel cross the U.S. border the most times, suggesting that these paths are most vulnerable to knowledge flow restrictions at the U.S. border. Notably, paths captured by the U.S. cross the U.S. border more times than those captured by any other region, with the exception of Israel.

Finally, Figure S10 shows information on the capture of outstanding paths by region. Figure S10A shows the number of captured outstanding paths, while Figure S10B shows the change in the number of outstanding paths.

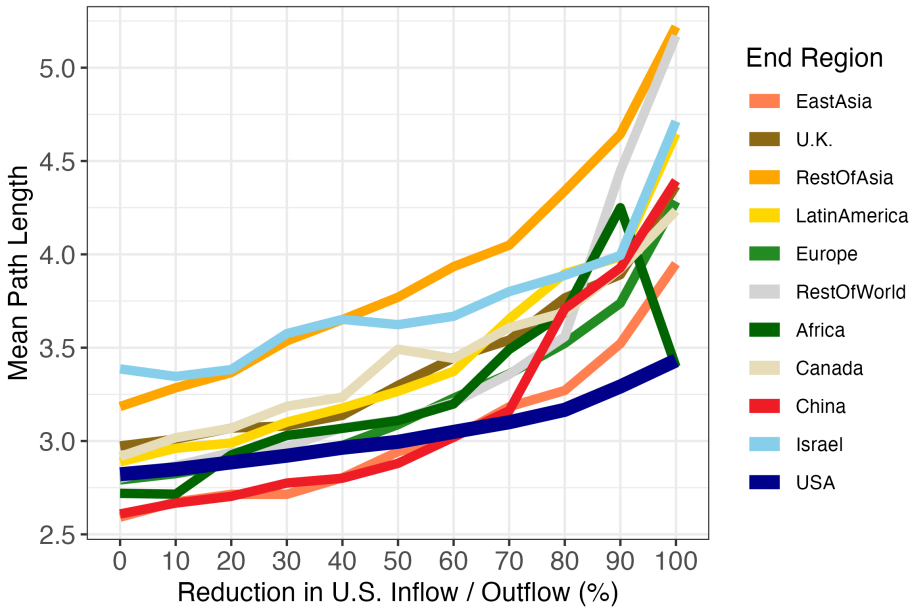


Figure S8: Mean Length of Realized Paths, for All Regions

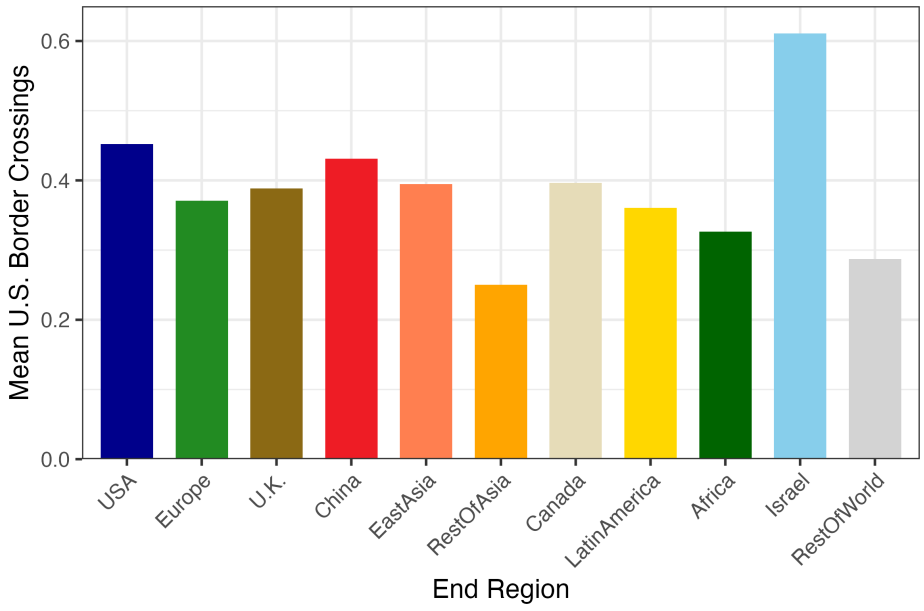


Figure S9: Number of Path Crossings of the U.S. Border by End Region

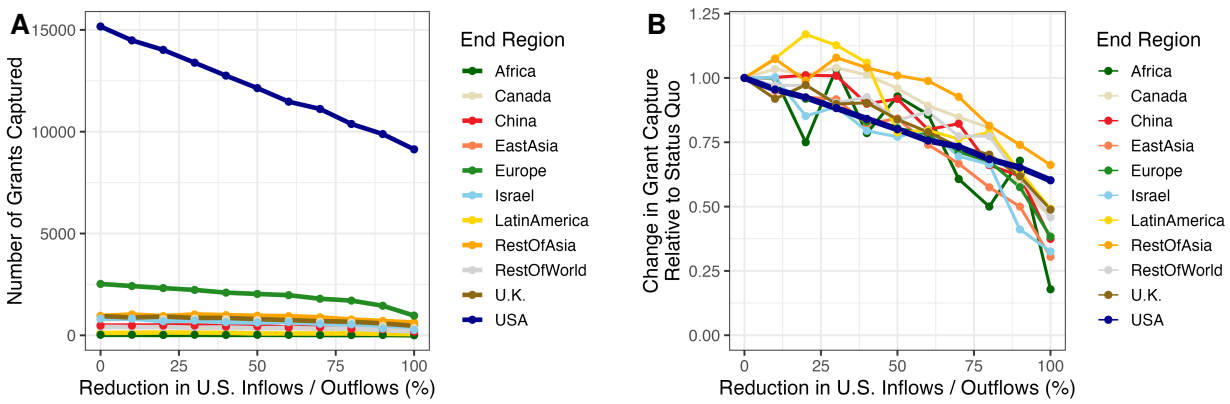


Figure S10: Outstanding Paths Captured by Region. (A) Number of paths captured by region. **(B)** Change in paths captured by region.