Frictions in the Flow of Academic Knowledge to Industry: Evidence from Simultaneous Discoveries

Michaël Bikard
London Business School
mbikard@london.edu

Matt Marx
MIT Sloan School of Management
mmarx@mit.edu

ABSTRACT

Abstract: Scientific discoveries in academia can spur innovation and growth, but only if that knowledge flows to relevant industry actors. One possible friction in the flow of academic knowledge to industry could be that many academic scientists are geographically isolated from firms performing R&D in relevant fields. But research at isolated institutions may be less applied, or simply of lower quality, confounding inference. We address the unobserved-quality problem by analyzing simultaneous discoveries where multiple researchers in different locations report the same finding in separate papers. We find that "twin" papers reporting a simultaneous discovery are 10-23% less likely to be referenced as prior art in firm-owned patents when not within commuting distance of a significant concentration of R&D activity in relevant fields. No effect is found for references from university-owned patents. Our results suggest that discoveries at isolated institutions may become orphaned, suggesting both implications for the science of science policy as well as firms' commercialization strategies.

Both authors contributed equally. We thank Jeff Furman, Christopher Liu, Fiona Murray, Scott Stern, Keyvan Vakili, and participants at the NBER Productivity Lunch and Harvard Business School TOM Alumni Research Workshop for useful comments. This work was supported by a Kauffman Junior Faculty Fellowship and be the Deloitte Institute of Innovation and Entrepreneurship at LBS.

Academic research is an essential engine of innovation and growth (Romer 1990; Grossman and Helpman 1993; Aghion, Dewatripont, and Stein 2008), with governments across the world investing billions[1] annually with the expectation that economic benefits will follow. Academic advances make new inventions possible because inventors can use the new knowledge as a guide in the invention process. One of the key economic benefits of academic research is therefore that it can be used by firms to increase R&D efficiency (Nelson 1959; Nelson 1982; Cohen, Nelson, and Walsh 2002). Thus, the economic impact of academic research depends critically on the flow of academic knowledge to industry.

Frictions in the flow of knowledge between academia and industry are liable to hinder innovation and economic growth. Even in the early 19th century, Charles Babbage highlighted the crucial role of science for "the arts and manufactures" and argued that the connection between science and those manufactures "should be rendered more intimate" so as to ensure that useful scientific discoveries get exploited by industry (Babbage 1832, 307). Indeed, inventors in firms do not always take advantage of useful academic advances. As Mokyr explains, scientific knowledge may "open doors hitherto closed" although "[o]pening such doors does not guarantee that anyone will choose to walk through them" (Mokyr 2002, 9). The ability of a society to create economic value from academic research depends not only on generating new knowledge, but also on the dissemination of that knowledge to firms. Yet, the circumstances under which academic knowledge flows or fails to flow to industry remain little understood.

Social scientists face a considerable challenge in assessing frictions in the flow of scientific knowledge from academia to industrial R&D labs. Firms might ignore academic discoveries because of these frictions, but they might also ignore them because those discoveries are less applied or more "fundamental"—i.e., not immediately useful for technology development. Geography, in particular, might play an important role in the exploitation of academic advances by firms.[2] The geographic distribution of industrial R&D rarely matches that of academic institutions, so academic scientists are therefore often geographically isolated from the corporate inventors that could use their research results. While we are interested in the marginal impact of the geographic environment on the dissemination of academic discoveries, a selection effect may confound our analysis if knowledge that is technologically more useful tends to emerge closer to industrial R&D activity. Certainly, recent evidence suggests that academic science is in part endogenous to local firms' research priorities (Sohn 2014).

---

[1] According to the Science and Engineering Indicators 2014, over $63bn was spent to fund research in US universities and colleges in 2011 up from $49bn in 2006 Source: http://www.nsf.gov/statistics/seind14/index.cfm/chapter-4/c4s1.htm#s3

[2] A large literature has highlighted that, conditional on the flow occurring, it is more likely to be reach firms that are located near academic institutions than by more remote ones (Jaffe, Trajtenberg, and Henderson 1993; Zucker, Darby, and Brewer 1998; Adams 2002; Furman and MacGarvie 2007; Belenzon and Schankerman 2013)

To address these selection issues, our paper uses a novel empirical strategy that exploits the common occurrence of simultaneous scientific discoveries (Merton 1961)[3]. When two or more discoverers submit their findings for publication at almost the same time, those papers disclosing the same discovery can be accepted, leading to the publication of "paper twins." By embodying the same piece of knowledge that emerged in multiple locations, these papers are a natural consequence of the duplication of effort in science and represent a potentially rich setting in which to study the impact of geographic isolation from industrial R&D activity on the dissemination of academic science. In practice, we measure the flow of academic discoveries to industry via references made by U.S. patents to 380 scientific publications disclosing 187 simultaneous discoveries. Our identification of geographic frictions in the flow of academic knowledge to industry is thus based on differential referencing of one "twin" paper reporting a scientific discovery in one location versus another twin reporting the same discovery in another location. We therefore extend a large literature that used references in patents and publications as a measure of knowledge flow (Jaffe, Trajtenberg, and Henderson 1993; Griliches 1998; Furman and Stern 2011; Galasso and Schankerman 2014). In particular, our empirical strategy presents three key advantages.

First, the use of patent references as a measure of knowledge flow is complicated by the possible existence of false positives as citations are often added ex-post for legal or strategic reasons (Alcácer, Gittelman, and Sampat 2009; Lampe 2012). This practice is linked to the doctrine of "Inequitable Conduct" which stipulates that patent applicants have a duty of candor and good faith when disclosing material prior art to the USPTO. In practice, the non-disclosure of relevant prior art might lead to patent invalidation, but it might also result in broader patents. Our identification strategy is facilitated by a key feature of the USPTO rules for recognizing prior art. Although inventors are required to disclose all relevant prior art, Rule 56 states that an inventor is not required to reference multiple sources disclosing the same prior art. Thus if a simultaneous discovery were relevant prior art, but the patent referenced only one of the "twin" papers reporting that discovery, not referencing another twin paper would not affect either the scope or the validity of the patent. In other words, our data are unique in that the references that we observe are unlikely to be driven by legal or strategic concerns.

Second, since academic discoveries tend to be published rather than patented, a focus on references to academic patents might lead to a considerable amount of false negatives. Studies using this measure (e.g., Jaffe, Trajtenberg, and Henderson 1993; Henderson, Jaffe, and Trajtenberg 1998) are

[3] Though our focus here is on the natural sciences, one should note that simultaneous discoveries also occur in the social sciences. The independent proofs of the existence of a competitive equilibrium in a market economy by McKenzie and by Arrow-Debreu in 1954 (Weintraub 2011) is only one famous example of this phenomenon in economic theory, and many other cases have been reported (Niehans 1995).

therefore likely to underestimate the influence of academic research on industrial R&D (Belenzon and Schankerman 2013; Roach and Cohen 2013). This paper gets around this difficulty by using non-patent references in patents. While scientific references from patents by no means capture every flow of academic knowledge to industrial R&D, Roach and Cohen (2013) report that they are perhaps the most reliable indicator.

Third, as is true of any case-control analysis, the reliability of inference depends critically on the quality of the controls. The difficulty in observing an appropriate control in the case of knowledge flow and patent citations is well known (for a critique of this issue, see Thompson and Fox-Kean 2005). Our ability to observe the same discovery being made in different contexts provides a rare opportunity to abstract away from the type of knowledge produced in different settings. In other words, the fact that we focus on the same discovery made simultaneously in multiple places means that we are able to observe patent references to academic papers that could have been made but were not.

Our results indicate that the geographic isolation of academic institutions reduces the flow of their discoveries to industry. Even after accounting for the effect of geographic distance, papers stemming from institutions not within 50 miles of relevant industrial R&D activity are much less likely to be referenced in a patent as prior art than are papers disclosing the same discovery but published at non-isolated institutions. In other words, the effect of geographic isolation does not result from distance alone. This result is robust to a number of different specifications, but it is only visible in references made by corporate inventors. In our data, academic inventors appear to reference papers produced at isolated and non-isolated institutions at similar rates. We therefore find evidence that scientific knowledge might flow more readily across areas in which firms conduct relevant R&D than outside of them. Thus, scientific discoveries made in geographically isolated institutions might end up entirely ignored by firms which might have benefited from exploiting them to develop new technologies.

These findings suggest that public investment in academic institutions that are geographically isolated from industrial R&D activity might fail to be exploited by industry inventors and therefore also fail to benefit the wider economy. Put another way, scientific discoveries at geographically isolated institutions may become "trees falling in the forest" without being heard. A second, and perhaps equally troubling implication of our findings is for the careers of scientists at geographically-isolated institutions. If our findings are correct, then two scientists of equal academic ability and with similar interest in having their work disseminated to the commercial world may have very different experiences simply by virtue of having been hired by an institution located near or far from firms conducting relevant R&D. Such processes could in turn lead to stratification of scientific careers and unequal scientific and financial rewards. Finally, our findings raise the question of whether firms might benefit disproportionately from paying closer attention to scientific research performed at geographically isolated academic institutions.

1. The Flow of Academic Science to Industry and Geographic Isolation

A. The Flow of Academic Science to Industry

Scientific knowledge can increase R&D efficiency because it guides the invention process. From the perspective of an inventor, R&D capability therefore depends on the person's knowledge (Nelson 1982). Joel Mokyr (2002) proposes a compelling account of the relationship between a society's knowledge, and its ability to invent. A society's understanding about nature and its regularities can help it invent, because the process of invention consists in a large part in using these regularities for a purpose. In line with this reasoning, large-scale empirical studies have established a link between university research and corporate patenting (Jaffe 1989) as well as productivity growth (Adams 1990). Surveys have provided illuminating insights about the exploitation of academic knowledge in industrial R&D. Mansfield (1998) studied large firms in seven industries and found that more than 5% of the total sales of those firms were directly due to innovations that could not have been possible without substantial delay in the absence of recent academic research. Cohen, Nelson, and Walsh (2002) used data from the Carnegie Mellon Survey and show that firms use academic knowledge both to generate new ideas and to address existing R&D problems. Overall, the impact of academic knowledge for industrial R&D appears to be large but highly heterogeneous across industries.

Despite having an institutional environment that fosters the dissemination of academic knowledge (Merton 1973; Dasgupta and David 1994; Stephan 1996), concerns have been voiced that frictions might prevent the dissemination of that knowledge to the commercial world. In the early 19th century, Charles Babbage argued that "the man of science should mix with the world" (Babbage 1832, 384) not only to ensure that he investigates important questions, but also so that knowledge flows to the eventual manufacturers. Babbage was worried that useful scientific discoveries might be ignored by firms even though they would benefit from exploiting them. Echoing Babbage's concern, NIH director Francis Collins recently declared that he was "frustrated to see how many of the [academic] discoveries that do look as though they have therapeutic implications are waiting for the pharmaceutical industry to follow through with them" (Harris 2011). Such inefficiency might exist if those firms have imperfect access to the newly produced knowledge. Mokyr (2002) proposed that "progress in exploiting the existing stock of knowledge will depend first and foremost on the efficiency and cost of *access* to knowledge." This paper explores this proposition empirically. In particular, we focus on the impact of the geographic isolation of academic research institutions.

B. Geography and the Dissemination of Academic Science

Geography might lead to frictions in the flow of academic science to firms because academic research institutions and industrial R&D labs are not always collocated. In their study of 8,074 commercial innovations introduced in 1982, Audretsch and Feldman (1996) find that the bulk of innovative activity in the US occurs on the coasts, and especially in California and in New England. Importantly, they find that innovative activity tends to cluster more in industries where knowledge spillovers play a decisive role. The sharp contrast between the geographic dispersion of academic scientists and the spatial concentration of industrial R&D activity is clearly visible in the case of the biotechnology industry. Audretsch and Stephan (1996) linked the location of biotechnology firms with that of the academic scientists who had relationships with these companies for the entire population of biotechnology firms that prepared an IPO in the early 1990's. While 69% of the firms in their samples were based in the Boston area, the San Francisco Bay area, and the San Diego area, those regions accounted for only 36% of all academic contacts. A number of academic institutions conducted leading edge academic research but did not appear to be connected to any local biotechnology firms. Such institutions include Yale University, the University of Michigan at Ann Arbor, the California Institute of Technology, the University of Alabama at Birmingham, Johns Hopkins University, UT Southwestern Medical Center, UC Davis, Pittsburgh University, Penn State University, etc.

The distinct geographic distribution of the suppliers of scientific knowledge in academic institutions and of the consumers of that knowledge in industrial R&D laboratories has fundamental implications in light of the localized nature of knowledge spillovers (Marshall 1895). Firms that are located at close proximity to academic institutions are known to benefit more from their research than firms that are located further away (Jaffe, Trajtenberg, and Henderson 1993; Zucker, Darby, and Brewer 1998; Furman and MacGarvie 2007; Belenzon and Schankerman 2013). By itself, the localization argument does not allow any prediction about the rate of dissemination of academic knowledge to firms. Yet, understanding *whether* the knowledge will flow as opposed to *who* benefits from that flow is especially important in light of the growing expectation that public investment in academic science translate into economic gains. The potential existence of frictions critically depends on the geographic distribution of the knowledge demand and supply, a topic that has been thoroughly studied by economic geographers (for a review, see Feldman & Kogler 2010). Some states or countries might experience very few such frictions because academic researchers are collocated with the relevant industry inventors. In other states, however, academic research institutions might be geographically isolated from the relevant industrial R&D activity, leading to considerable frictions in the dissemination of academic knowledge. Taken together, the literature on the localization of knowledge spillovers and that on the economic geography of science and technology naturally lead to our prediction that the geographic isolation of

many academic research institutions from the relevant industrial R&D activity creates frictions in the flow of academic knowledge to firms.

C. The Geographic Isolation of Academic Research Institutions: an Overview

The tremendous potential of academic research for regional economies is often illustrated by description of highly visible successes of MIT in the Boston/Cambridge area or Stanford University in the Silicon Valley (e.g., Jaffe 1989). Yet, unlike MIT and Stanford, many academic research institutions are not located near major hubs of industrial R&D. In fact, the concentration of high technology industries and the dispersion of academic research institutions entails that the large majority of those institutions are geographically isolated from relevant inventors in industry.

Figure 1 about here

Figure 1 shows the geographic isolation of academic research institutions from commercial R&D in the U.S. All institutions shown published at least two papers in the top 15 scientific journals between 2000 and 2010.[4] On this map, we label a campus as geographically isolated (blue pin) if fewer than 5000 patents were awarded to inventors living within a 25-mile radius. The 10 most isolated yet productive institutions are labeled on the map.[5] Using these measures, the University of Wisconsin at Madison and the University of Illinois at Urbana-Champaign are the most productive academic research institutions that are geographically isolated from industrial R&D. Overall, over 90 top US academic research institutions appear geographically isolated from industrial inventors. They include prestigious institutions such as Dartmouth College in Hanover, NH and Cornell University in Ithaca, NY. Of course, in reality, geographic isolation from relevant industrial R&D activity varies across fields and overtime. Our empirical analysis will therefore measure isolation in a more refined way which accounts for this variance.

One approach to examine the frictions stemming from the geographic isolation of academic institutions would be to simply compare patent references to papers from isolated vs. non-isolated institutions. Indeed, unreported results indicate that the institutions on this map that are more isolated are less likely to see their discoveries referenced by commercial inventors. However, the challenge in interpreting the results from such an analysis would be considerable. Lower rates of patent referencing publications from isolated institutions might be driven by frictions in the dissemination of knowledge.

---

[4] These are Nature, Science, Cell, New England Journal of Medicine, Journal of the American Medical Association, Lancet, CA: A Cancer Journal for Clinicians, Nature Genetics, Nature Materials, Nature Medicine, Nature Immunology, Nature Nanotechnology, Nature Biotechnology, Cancer Cell, and Cell Stem Cell. Impact factor was measured as a five-year rolling average between 2005 and 2009.

[5] Productivity is measured by counting the number of papers published by each institution in the top-15 impact factors journals between 2000 and 2010.

However, the same negative relationship might also stem from the fact that geographically-isolated academic institutions conduct research of a more basic nature, or of lower quality. Perhaps worse scientists or those more oriented toward basic research sort into institutions that happen to be geographically isolated, or perhaps the institutional environment has endogenously evolved to be less supportive of science and its dissemination. After all, "industrial activity, especially, but not only, in high tech sectors, provides unique observational platforms from which to observe unusual classes of natural phenomena" (Rosenberg 1994, 141). In line with this argument, recent research confirms the popular intuition that industrial R&D units can have a considerable impact in shaping the research conducted by nearby universities (Sohn 2014).

These challenges make empirical analysis of the frictions stemming from geographic isolation even more elusive. The remainder of the paper is dedicated to establishing a causal link between the geographic isolation of academic institutions from firms' R&D activity and the failure of scientific discoveries to be exploited by industry inventors. The perfect experiment would require us to have identical discoveries made by identical scientists working at institutions that are identical except for their geographic location, an arrangement we are highly unlikely to find. However, we can make progress by measuring differences in the likelihood that simultaneous discoveries in different geographic locations—some isolated from firms' R&D—will disseminate to industrial R&D. The next section describes how we find simultaneous discoveries and leverage them to identify geographic isolation as a friction in the flow of knowledge from academia to industry.

2. Data Construction

A. Illustration

Our identification strategy hinges on simultaneous scientific discoveries. Before describing the process by which these were found, we illustrate the nature of a simultaneous discovery with an example. The August 1998 issue of *Cell* contains two papers reporting the same scientific discovery.

> **Cleavage of BID by Caspase 8 Mediates the Mitochondrial Damage in the Fas Pathway of Apoptosis**
> *Li, Zhu, Xu, and Yuan at Harvard Medical School, Boston MA*
> We report here that BID, a BH3 domain-containing proapoptotic Bcl2 family member, is a specific proximal substrate of Casp8 in the Fas apoptotic signaling pathway. While full-length BID is localized in cytosol, truncated BID (tBID) translocates to mitochondria and thus transduces apoptotic signals from cytoplasmic membrane to mitochondria. tBID induces first the clustering of mitochondria around the nuclei and release of cytochrome c independent of caspase activity, and then the loss of mitochondrial membrance potential, cell shrinkage, and nuclear condensation in a caspase-dependent fashion. Coexpression of BcixL inhibits all the apoptotic changes induced by tBID. Our results indicate that BID is a mediator of mitochondrial damage induced by Casp8.

Bid, a Bcl2 Interacting Protein, Mediates Cytochrome c Release from Mitochondria in Response to Activation of Cell Surface Death Receptors
Luo, Budihardjo, Zou, Slaughter, and Wang at the University of Texas Southwestern Medical Center, Dallas TX
We report here the purification of a cytosolic protein that induces cytochrome c release from mitochondria in response to caspase-8, the apical caspace activated by cell suface death receptors such as Fas and TNF. Peptide mass fingerprinting identified this protein as Bid, a BH3 domain-containing protein know to interact with both Bcl2 and Bax. Caspase-8 cleaves Bid, and the COOH-terminal part translocates to mitochondria where it triggers cytochrome c release. Immuno-depletion of Bid from cell extracts eliminated the cytochrome c releasing activity. The cytochrome c releasing activity of Bid was antagonized by Bcl2. A mutation at the BH3 domain diminished its cytochrome c releasing activity. Bid, therefore, relays an apoptotic signal from the cell surface to mitochondria.

Both papers report the discovery of an important molecule involved in the cell death or apoptosis. The two teams found that after activation of the death receptors on the cell membrane, the death signal is carried to the mitochondria by a cytosolic protein called BID. Confirming that these two papers truly report the same scientific discovery, an August 21 2000 article in *The Scientist* notes that "[t]hese two *Cell* papers outline two independent identifications of a critical missing link in [the apoptosis] signaling pathway" (Halim 2000). As occurs frequently in the case of simultaneous discoveries, both papers were published back-to-back (pages 481-490 and 491-501) in the same issue of the same journal. As noted below, editors receiving manuscripts that report the same (or very similar) findings around the same time frequently elect to publish them back-to-back in order to underscore the reliability of important discoveries.

We exploit simultaneous discoveries to overcome the aforementioned quality-of-the-discovery problem inherent to analyzing differential rates of knowledge flow. Specifically, the knowledge disclosed in each "twin" paper reporting the simultaneous discovery should have a similar risk of being exploited by industry inventors. In practice, we are able to measure the rate of dissemination by tracking the references to each scientific paper in patents. Of the two papers reporting the BID protein, the paper located in Boston, where local firms perform R&D in similar fields, received more references from patents than did the paper in Dallas, which is largely isolated from relevant industry. In the following section, we describe how we find simultaneous discoveries.

B. Identifying Simultaneous Discoveries

The data for this study is based on the first automatically and systematically collected dataset of simultaneous discoveries. The full dataset consists of 1,246 papers, published between 1970 and 2009, disclosing 578 simultaneous discoveries. The algorithm that was built to identify those paper twins scrolls

through the scientific literature to identify instances in which two papers are consistently cited in the same parenthesis, or adjacently.

The method is detailed in a companion paper (Bikard 2012) but for convenience, its main principles are described here. The algorithm is rooted in the results from two distinct literatures. On the one hand, sociologists of science have found that citations provide a window into the scientific community's allocation of credit. In a sense, the community uses citations as a "vote" regarding which team deserves the credit for a given discovery (Cozzens 1989). As a result, systematic co-citation in the scientific literature indicates that the community has decided that the credit for a specific discovery ought to be shared across different teams. While occasional co-citation might point to discoveries that are complementary rather than simultaneous, systematic co-citation indicates that two of more papers share the credit for the same discovery. On the other hand, citations provide a convenient similarity metric to relate documents (Marshakova 1973; Small 1973). As such, they can be used to map science, but can also be fed into search engines pointing to related papers. As an example, as CiteSeer uses co-citations to compute the relatedness between academic papers (Giles, Bollacker, and Lawrence 1998). Recent studies have suggested that these algorithms can be made even more precise by considering citation proximity within each paper. For instance, papers that are co-cited in the same sentence tend to be particularly similar to each other (Gipp and Beel 2009; Tran et al. 2009). The algorithm that was used here goes one step further and considers pairs of scientific publications that are consistently cited together—i.e., in the same parenthesis, or adjacently.

In practice, the algorithm uses five steps. In step 1, a dataset consisting of information about 42,106 scientific articles was built using ISI Web of Knowledge. It is composed of all the non-review research publications that appeared in the 15 scientific journals having the highest impact factor between 2000 and 2010. In step 2, each reference in all of these articles were given a unique identifier using Pubmed and CrossRef. Of 1,294,357 references, 744,583 unique references were identified. Step 3 generates a database of pairs of all references that were (a) co-cited at least once, (b) written no more than a calendar year apart, (c) have no overlapping authors, (d) in which at least 5 citations for each reference are observed in the dataset of 42,106 citing articles. Of the 17,050,914 pairs of papers that were considered, 449,417 pairs meet these criteria. Step 4, consists in establishing a first measure of co-citation. A Jaccard co-citation coefficient was used following the scientometric literature. It consists in the intersection over the union of citations that both papers receive for each pair. 2,320 pairs of papers were selected that had a co-citation coefficient superior to 50%. Finally, step 5 consists in selecting those pairs for which 100% of the co-citations took place in the same parenthesis or adjacently. To do so, a parsing algorithm examined all the co-citing articles. 495 pairs for which fewer than 3 co-citing articles could be parsed were excluded. Of the remaining 1,825 pairs, 720 had been cited adjacently in 100% of the co-

citing articles. These 720 pairings of 1,246 papers disclose 578 unique discoveries since there are instances of discoveries involving three or more teams.

The extent to which the resulting pairs are actually instances of simultaneous discoveries was tested in several ways (Bikard 2012). First, paper twins should be published around the same time. As noted above, the algorithm matches on co-citation and not on publication month. If two alleged paper twins were not really disclosing the same discovery, one would expect them to be on average six months apart or more.[6] The 720 paper twins in the entire dataset were in fact published on average 1.8 months apart, a lag considerably shorter than the average time between paper submission and publication. In fact, 373 pairs of twins were published the exact same month, and 267 of them were published in the same issue of the same journal.[7] Second, the Pubmed related citation algorithm uses semantic similarity to match scientific papers. Since the large majority of the 1,246 papers also appear in Pubmed, we can use this algorithm to measure the semantic similarity between pairs of papers that our algorithm identified as disclosing the same discovery. If the pairs were not very closely related, they should not be using the same words and should therefore be ranked far from each other. Pubmed ranks two papers of the same pair right next to each other 42% of the time. The rank difference is inferior to 10 for 90% of the pairs.[8] Third, 27 scientists who had been corresponding authors on at least one of the 1,246 papers were interviewed. Importantly, none of them contested the fact that they were sharing the credit with another team for the same discovery and some were bitter about it.[9] Five of the interviewees claimed that their idea had been stolen by the other team. Confirming that the algorithm uses very conservative criteria, the

[6] The algorithm does not match on month, but it limits the consideration set of papers to pairs that were published no more than a calendar year apart (we considered that papers published more than 23 months apart cannot be disclosing the same discovery). This choice is limiting because many independent discoveries are known to have taken place years apart of each other (see Ogburn and Thomas (1922) for numerous examples). However, since credit for scientific discoveries is a function of priority, it is reassuring that we ended up with pairs of papers published very close to each other. Besides, for our study, it is important that the paper emerge around the same time so they have the same chance of being used by corporate inventors.

[7] When two teams send manuscripts to the same journal describing essentially the same findings around the same time, editors sometimes decide to publish them back-to-back, therefore recognizing a tie in the race for priority, and allowing both teams to receive equal credit for their work. Well-known examples of back-to-back publications include that of evolution by natural selection by Darwin and Wallace in the *Journal of the Proceedings of the Linnean Society of London* published on 20 August 1858 and the discovery by Richter and Ting of the J/ψ meson published in *Physical Review Letters* on 2 December 1974. While simultaneous discoveries appear often (but not always) back-to-back in scientific journals, one should note that not every back-to-back publications correspond to simultaneous discoveries (Drahl 2014).

[8] Rank difference calculated after dropping articles that are published more than a calendar year apart. For more information about the the Pubmed related citation algorithm, see http://ii.nlm.nih.gov/MTI/Details/related.shtml

[9] Sharing the credit does not mean that the two (or more) papers were identical. Two scientific articles written by two different teams are never completely identical, and differences might exist in the tools/methods used, in the number of experiments, or in the interpretation of the results. However, the fact that the papers share the credit indicates that the scientific community considers that both teams provided convincing evidence to support their claim of priority in making the discovery.

interviewees also revealed in several cases that more teams than we were aware of had claimed to have taken part in the simultaneous discovery. One should keep in mind that, by design, our algorithm excludes any priority claim that is not clearly visible through the citations of the broader scientific community.

C. Measuring the Flow of Academic Science to Industry

Tracking the flow of academic science to industry empirically is challenging because such flows can take a variety of forms. In a landmark paper, Jaffe, Henderson and Trajtenberg (1993) proposed that patent citations can be used to measure knowledge flow. In so doing, they laid the foundation for a rich literature that has provided fascinating insights about the dissemination of knowledge (e.g., Henderson, Jaffe, and Trajtenberg 1998; MacGarvie 2006; Singh and Agrawal 2011; Galasso and Schankerman 2014). Patent citations are a tremendously useful measure, both because they are systematic and because they are readily available. This measure is not without important limitations, however. First, patent citations have legal implications since they delimit the scope of an invention. The doctrine of "Inequitable Conduct" means that omission of information material to patentability can lead to the invalidation of the patent. Patent citations are therefore often added by patent attorney and patent examiners (Alcácer and Gittelman 2006; Alcácer, Gittelman, and Sampat 2009) and they can be used strategically (Lampe 2012). In other words, inventors, attorneys and examiners have an incentive to add citations to knowledge that was not used for invention, therefore significantly complicating the task of the empiricist using citations as a measure of knowledge flow. Second, each patent is by definition unique, making interpretation of non-citation difficult. Concerns regarding the definition of a control group of non-citing patents has led to major debates in the literature studying the localization of knowledge spillovers (Thompson and Fox-Kean 2005; Henderson, Jaffe, and Trajtenberg 2005). Third, knowledge is often not patented and therefore not observable using this type of measure (Griliches 1990). This concern is particularly salient in the case of the knowledge produced by academic institutions because those institutions primarily disclose knowledge through scientific publications rather than patenting (Ajay Agrawal and Henderson 2002; Belenzon and Schankerman 2013; Roach and Cohen 2013).

Characteristics of our empirical setting allow us to largely address all three challenges. First, there is no legal requirement to refer to every paper disclosing the same simultaneous discovery. According to USPTO Rule 56 (37 CFR 1.56): "information is material to patentability when it is not cumulative to information already of record or being made of record in the application." In other words, if multiple papers disclose the same knowledge, referring to one of the papers is sufficient. References in our setting are therefore much less likely to be driven by legal or strategic considerations. Second, while every patent is by definition unique, the same is not true for scientific publications. The patent system does not recognize "ties" in the race for priority and simultaneous or independent inventions are therefore

the topic of important debates in the legal literature (Vermont 2006; Lemley 2007). The same is not true in science. As we described above, when two papers make the same discovery and send it for publication at around the same time, multiple papers can be published disclosing very similar knowledge (e.g., Cozzens 1989). Third, since most academic knowledge gets published rather than patented, we focus on nonpatent references in patents. Roach and Cohen (2013) study the validity of citations as a measure of knowledge flow from public research and warn that they cannot account for private interactions such as consulting, cooperative research ventures, and contract R&D in which knowledge does not get codified in the form of papers or patents.[10] However, they emphasize that "in all our analyses, we find that citations to *nonpatent* references, such as scientific journal articles, correspond more closely to managers' reports of the use of public research than do the more commonly employed citations to *patent* references" (Roach and Cohen 2013, 505). In addition, the fact that the large majority of simultaneous discoveries in our sample belong to the life sciences is advantageous since this is a field in which the use of publications and patents by firms is particularly widespread, making scientific references from patents (but not to other patents) a more accurate indicator of knowledge flow than they might be in other specialties (Roach and Cohen 2013).

The simultaneous discoveries we focus on are highly visible to the scientific community for three reasons. First, the discoveries in our data attracted the attention of several teams of scientists. Second, editors of scientific journals have found the discovery important enough to collectively publish more than one paper disclosing that discovery. Third, we identified the "paper twins" by focusing on systematic co-citation. This means that poorly cited simultaneous discoveries would not enter our dataset. That these simultaneous discoveries are highly visible is likely to mean that our empirical test is conservative and that our results might understate the differential in knowledge spillovers in the case of (less visible) non-simultaneous discoveries.[11]

Unlike patent citations to other patents, patent references to scientific publications are not readily available through existing databases. Each patent contains a list of non-patent references in the "Other References" section (named "sciref file" in the Dataverse database (Li et al. 2014)), which are provided as unstructured text strings. One might consider searching for the title of the paper and journal among the scientific references listed in the patent, but from our initial attempts to do so we found too many variations to avoid myriad Type I errors. We found frequent abbreviations of words within the title and

[10] The fact that references cannot measure this type of private and uncodified knowledge flow is likely to bias our results toward under-estimating the impact of isolation as a driver of frictions because these private interactions are more likely to be local (Audretsch and Stephan 1996; Zucker, Darby, and Brewer 1998)

[11] A related study might investigate the impact of geographic isolation on the dissemination of scientific knowledge to other scientists. Unfortunately, since we identify simultaneous discoveries based on systematic co-citations in the academic literature, our dataset is not adapted to address this question.

journal name, substantial truncations of titles, and occasional misspellings. Instead, we elected to use four more easily matched criteria: 1) the surname of the first author, 2) the year of the journal, 3) volume number of journal, and 4) the starting page number. This tuple is highly unlikely to be non-unique; in order for this to occur, two authors with the same surname would have had to publish articles in different journals that had the same volume number in the same year; moreover both articles would have to start on the same page.

We begin by parsing the first author's name, year, volume, and first page from the scientific references listed in the patent. A similar exercise is performed for the scientific papers, and then the two groups of {author surname, year of journal, journal volume, initial page number} characteristics are matched with each other. We use the matches produced from these four criteria as a first pass to create a superset of possible matches and then inspect those by hand for Type II errors. This exercise produces the dependent variables used in this paper. At the paper level, *NUMREFS* counts the number of references from any patent by 2010. Our main analysis, however, focuses on paper-patent dyads where we predict whether a given dyad will or will not be linked by a reference. For dyad analysis, the dependent variable is *REFERENCED*.

D. Linking Simultaneous Discoveries and Patents

We apply the above methodology to our 578 simultaneous discoveries published in 1,246 papers. Given our interest in the flow of academic research to industry, we drop references from patents assigned to universities or other academic entities. We then eliminate self-references in two ways. First, if the surname and first initial of any author on the paper matches any inventor on the patent, we remove the paper-patent dyad from consideration. (Note that references from within the same organization, typically excluded from patent-citation studies, are not of concern in our research design because the patents are from firms while the papers are from academic institutions.) Second, and perhaps more importantly, we manually reviewed the acknowledgments section of each "twin" paper and then removed paper-patent dyads where the patent assignee was acknowledged in the paper as a sponsor of that research. Applying these restrictions reduces the sample to 380 papers reporting 187 simultaneous discoveries, with 1,910 scientific references from 1,281 patents.

As the objective of our identification strategy is to compare the likelihood of differently-located simultaneous discoveries having been referenced by patents, we then construct a dataset where every patent that referenced any of our 380 papers is paired with all papers that disclose that same simultaneous discovery. For example, given a pair of "twin" papers where one of the papers is referenced by a later patent, we also create an observation for that same patent together with the twin paper that was not

referenced but could have been, given that the twin papers disclose the same simultaneous discovery. Figure 2 illustrates the setup.

Figure 2 about here

For each paper-patent dyad representing a (potential) scientific reference, we calculate several characteristics of the dyad including the time lag between the publication of the paper and the potentially referencing patent (*TIME_LAG)*, the geographic distance between them (*DIST*, or dummies indicating a particular distance range), and whether the paper and patent are in the same country (*SAME_COUNTRY)* and city (*SAME_CITY)*. If both the paper and the (potentially) referencing patent are both located in the U.S., we also calculate whether they are in the same Metropolitan Statistical Area (*SAME_MSA*). For North America, we calculate (*SAME_STATE)* using Canadian provinces and U.S. states.

We also include the paper-level variables from Table 1 to account for characteristics of the individuals and their institutions involved with the simultaneous discovery. At the level of individual scientists, less commercially inclined researchers might take jobs at institutions that happen to be isolated from industrial R&D. We consider that patenting of the simultaneous discovery may offer a window into researchers'' proclivities to actively disseminate their knowledge to firms beyond publishing a paper. *PAPER_PATENTED* indicates whether one of the authors of the focal paper patented the discovery in question, forming a "patent paper pair" (Murray 2002). An algorithm was built to find whether each of the papers in the dataset has a patent pair.[12] At the level of the academic institutions, the establishment of a technology transfer office (TTO) may indicate a commitment to having scientific discoveries exploited by industry (Jensen and Thursby 2001; Hellmann 2007). The *INSTITUTION_TTO* variable therefore indicates whether the focal institution had established a technology transfer office before the paper was published.[13] Variable definitions are in Table 1; descriptive statistics are in Table 2.

Tables 1 and 2 about here

E. Measuring Geographic Isolation

To measure geographic isolation from relevant industrial R&D, we focus on inventive activity (a) in the relevant field (b) within 5 years of the discovery and (c) within a specific radius of the institution. Isolation is operationalized as follows. We start by collecting the technological subclassifications from all patents, whether industrial or academic, that contain scientific references to one of the 380 "twin" papers

[12] For each article, the algorithm finds the patents which (a) were filed on the year of paper publication or the year preceding it (b) list at least 2 authors of the paper as inventors and (c) list as assignee at least one organization employing the paper's authors. The details of this algorithm are detailed in a companion paper (Bikard 2012).

[13] *INSTITUTION_TTO* has fewer observations than other variables because it is defined only for U.S.-based universities. Most analyses include also non-university academic institutions such as research institutes.

in our sample in order to have the most complete possible representation of USPTO patent subclasses that are applicable to the simultaneous discoveries.[14] Patents referencing the papers that report the simultaneous discoveries are categorized into 712 unique subclasses. For each subclass, we then collect all non-university patents belonging to that subclass, whether or not they reference any of the twins in our study. We find a total of 1,430,822 corporate patents that were categorized by the USPTO into one of the 712 technology subclasses.

We then construct "hubs" of industrial R&D activity as follows. For each of the 712 technology subclasses that characterize our simultaneous discoveries, we collect the locations in which those non-university patents are found in that subclass. For each location, we count the number of patents in that same subclass within a 50-mile radius for each half-decade. We divide those two figures to yield the percentage of overall patenting activity from that technology subclass occurring in that location. We label a location as a "hub" of industrial R&D for that subclass if more than 5% of patents in that technology subclass are located within a 50-mile radius. Because this threshold can easily be exceeded in technology subclasses with very few patents (e.g., in a subclass with only 20 patents, every location has at least 5% of patenting), we require that a location have at least five patents in that subclass to qualify as a "hub." This exercise yields a list of R&D hubs for each of the 712 technology subclasses relevant to our simultaneous discoveries within five years of the publication date. (Some subclasses are widely distributed across locations and thus do not have any hubs.)

To determine whether a given academic paper is isolated from relevant industrial R&D, we first make a list of the technological subclasses for all patents that referenced either the focal paper or any of its twins. These patent subclasses delimit the relevant scope of R&D activity for that simultaneous discovery. For each twin paper reporting that simultaneous discovery, we then check whether there is at least one R&D hub within 50 miles (i.e., commuting distance) of the corresponding address of the focal paper (likely the location of the lab where the research was conducted). If we cannot find a hub within 50 miles, we set *ISOLATED* to 1 for that paper.

Important to note is that isolation is a paper-level attribute, neither an institution- nor city-level attribute. Either an institution or a city may be isolated from one field but close to hubs of R&D for others. For example, Dallas is isolated from biotechnology but close to semiconductor R&D; the opposite is true for Boston. It is also possible that the concentration of R&D shifts over time, which motivates our use of five-year windows.

To illustrate the concept of isolation from relevant industrial R&D, we return to our simultaneous discovery from above. Again, we examine two papers in the August 1998 issue of *Cell*, one at Harvard

[14] We began by using top-level classifications but found these too broad, with many papers lumped into the same classification which contained tens of thousands of patents.

Medical School in Boston, MA and another at the University of Texas - Dallas. In determining whether either of these research teams was commercially isolated, we first note that 19 patents (firm-owned or university-owned) listed one of these papers as a scientific reference.[15] We then define the scope of relevant R&D by obtaining the USPTO technological subclassifications for these patents. A few have the same classification, yielding 17 subclasses: 424:187; 434:243,325,375,4,7; 514:12,210,21015,34,44,44R,45; 530:300,326; 536:231; and 540:355.

The next step is to locate "hubs" of industry R&D in these technological areas. We find 3858 firm-owned patents that were assigned to these subclasses during 1995-1999 (again, the article was published in 1998). The locations with R&D "hubs" containing at least five patents and more than 5% of patenting activity for the above 17 subclasses include Milan, Italy; La Jolla, Santa Clara, and Solana Beach, California; Canton, Lexington, and Weston, Massachusetts; Chevy Chase and Silver Spring, Maryland; Berkeley Heights, Old Bridge, and Teaneck, New Jersey, and Bainbridge Island, Washington.

Having constructed the set of non-isolated cities for the technological subclasses corresponding to those patents, we then check whether either of the twin papers is within commuting distance (i.e., 50 miles) of any of those. While Harvard Medical School in Boston is within commuting distance of Canton, Lexington, and Weston, Massachusetts, Dallas is far from any of the cities listed above as having at least five patents and more than 5% of patents in the subclass. Thus we classify the paper published in Dallas as isolated from relevant industrial R&D.

Applying this definition of isolation, 72.4% of our paper twins are isolated from relevant industrial R&D. Our definition of isolation is somewhat conservative, requiring only 5% of patents in the relevant subclass. Less conservative formulations (e.g., requiring more than 10% of patenting in the subclass), yield similar results and label close to 90% of papers as isolated.

F. Empirical Setup

We examine the impact of the geographic isolation of academic institutions on the flow of public research to industrial R&D by examining the references of academic papers disclosing the same discovery in corporate patents. An observation is a dyad of a published paper reporting a (simultaneous) discovery and a patent that is at risk of referencing the paper as non-patent reference. Our analysis leverages the simultaneous-discovery nature of our data since a patent that references one paper is presumably at a similar risk of referencing any of its "twins" as described in Figure 2. We specify a linear probability

[15] Both papers were referenced by three patents (7452869, 7638324, and 7745109). The Dallas paper was referenced exclusively by three patents: 6503754, 7247700, and 7829662. The Boston paper The Boston paper was referenced exclusively by 13 patents: 6221355, 6245885, 6326354, 6645501, 6692927, 6773911, 6946458, 7026472, 7371834, 7381713, 7514413, 7635693, and 7772202.

model with fixed effects for each group of paper twins reporting a simultaneous discovery, but our results are robust to a conditional logit specification. The regression equation is given as

$$REFERENCED_{kij} = \alpha ISOLATED_{kij} + \gamma_j + \delta X_{ki} + \varepsilon_{kij}$$

where *j* represents the simultaneous discovery, *i* represents the paper reporting the simultaneous discovery, and *k* represents the potentially-referencing patent. $ISOLATED_{kij}$ is our main explanatory variable and is defined at the paper-patent dyad level as described in section 2.E. γ_j is our simultaneous discovery-level fixed effect. For this linear equation to identify the average effect of geographic isolation on the flow of public research to industry, we implicitly assume that the potential variance in within simultaneous discovery paper quality is orthogonal to the location of that paper's institution. Finally, X_{kj} is a vector of covariates including the geographic distance between the paper and patent. Some specifications also include city and academic institution-level fixed effects. Standard errors are clustered at the level of the simultaneous discovery.

3. Empirical Results

Table 3 presents the basic results of our analysis. Column (1) presents a simple cross section without fixed effects for simultaneous discoveries. The dependent variable *NUMREFS* counts the number of times each paper is referenced by a patent. As one might expect, papers where one of the authors has patented the discovery (*PAPER_PATENTED*) accrue considerably more references from firms conducting R&D. For journal impact factor (*PAPER_JIF*), positive effects appear to accrue only at the very high end of the distribution. Interestingly, the paper's institution record of publishing in the top 15 scientific journals does not appear to influence the paper's ability to accrue references, as shown by the lack of statistical significance on *INSTITUTION_PRESTIGE.* Finally, papers that are geographically isolated from industrial R&D in relevant technological subclasses (*ISOLATED*) receive about ten fewer patent references than those that are not isolated.

Table 3 about here

Although the count model of column (1) suggests that isolated discoveries are considerably less likely to be referenced by patents, without fixed effects for the simultaneous discovery this result is vulnerable to criticism. The negative relationship between isolation and the number of references might arise not from frictions in knowledge flow but instead result from the endogenous sorting of research projects in institutions that are located close to relevant industrial R&D activity. For the remaining models, we shift to an analysis of dyads formed by patents referencing one of the papers that reports a particular simultaneous discovery, with one observation for each of the "twin" papers it either did reference or might have referenced, as depicted in Figure 2. The dependent variable *REFERENCED*

indicates whether or not the patent referenced the paper in the dyad. Here, we are able to account for the temporal separation between the paper's publication and the granting of the patent (*TIME_LAG*), as well as the spatial separation between the corresponding address of the paper and the primary inventor on the patent (*DIST*, other *DIST* dummies), correcting for the curvature of the earth.

These dyad analyses begin in column (2). As one might expect, U.S.-based papers are somewhat more likely to be referenced by USPTO patents (*PAPER_US*). As in column (1), whether one of the paper's authors patented the discovery (*PAPER_PATENTED*) strongly affects the likelihood of being referenced, possibly an indicator of commercially oriented efforts on behalf of the scientists or their institution. Journal impact factor (*PAPER_JIF*) is not impactful among our articles, perhaps because we sample from the 15 highest impact-factor scientific journals. Perhaps surprisingly, but consistent with column (1), papers from institutions with a track record of publishing in the top 15 scientific journals (*INSTITUTION_PRESTIGE)* do not appear more likely to accrue references in patents. Moreover, we add the dyad-level covariate corresponding to the lag between the publication of the focal paper and the application date of the potentially-citing patent (*TIME_LAG*), which does not appear consequential. As shown by the negative and statistically significant coefficient on *ISOLATED*, papers from institutions that are isolated from the relevant industrial R&D activity are less likely to be referenced by a patent referencing that simultaneous discovery.

Still, the negative coefficient on *ISOLATED* might be attributable largely to spatial separation (Jaffe, 1989; Adams, 1990; Audretsch & Feldman, 1996; Belenzon & Schankerman, 2013). In other words, papers that are geographically isolated from relevant industrial R&D activity may simply be further away from potentially-citing patents in our paper-patent dyads and thus less likely to be referenced by patents that reference the simultaneous discovery. In Column (3) we add a control for the (logged) spatial distance separating the focal paper and potentially-citing patent. Not only do we see little impact of spatial separation between the paper and patent—although the coefficient on *DIST* is negative, as one might expect, its statistical significance fails to reach even the 10% level—but the magnitude and statistical significance of the coefficient on *ISOLATED* is largely unaffected by controlling for distance. Similar results are recovered in column (5) when including various indicators for levels of distance as opposed to a single, continuous variable as the attenuating effect of distance on diffusion is unlikely to be linear (Singh and Marx 2013).[16] The coefficient on *DIST1-10* miles is both positive and has strong statistical significance, suggesting that patents filed within ten miles of the focal paper are more likely to reference them than other patents. Thus scientific discoveries are more likely to be noticed by commercial

[16] Results are also robust to a more finely crafted set of distance dummies: 1-10 miles, 11-20, 31-40, 41-50, 51-75, 76-100, 101-150, 151-200, 201-300, 301-400, 401-500, 501-750, 751-1000, 1001-1500, 1501-2000, 2001-2500, 2501-4000, and 4001-6000. The omitted category is greater than 6000 miles.

inventors who are very closely collocated, in contrast to previous literature suggesting that spillovers are localized at the MSA, state, and even country level. That the *ISOLATED* covariate is negative and significant in both columns (4) and (5) suggests that isolation from relevant industrial R&D plays an independent role in shaping knowledge flow.

In Table 4 we introduce fixed effects for the institution and location of the academic paper to account for the possibility that our results might be driven by unobserved characteristics of the publishing institutions themselves or the cities in which they are based. In column (1) of Table 4 we add fixed effects for each publishing institution (i.e., the corresponding address of the publication, which is likely the location of the laboratory where the research was conducted), which drops the institutional-prestige variables. The coefficient on *ISOLATION* is negative and statistically significant, corresponding to a 9.7% lower likelihood of a focal academic paper being referenced by an industrial patent. This analysis suggests that our results are not driven by unobserved characteristics of academic institutions such as a (difficult to observe) culture oriented toward commercialization. Rather, the likelihood of being referenced by relevant patents differs for papers from the same institution depending on the composition of the local industrial R&D activity. For example, the University of Texas – Dallas is isolated with respect to certain areas of biomedical R&D but is one of the hotbeds of semiconductor R&D. The result in column (4) of Table 4 suggests that the connection between geographic isolation and knowledge flow from academia to industry is not due simply to at a few highly-referenced, non-isolated institutions.

Table 4 about here

In column (5) we instead apply city fixed effects in order to assess whether our results are driven by unobserved characteristics of cities such as a culture that promotes the exchange of knowledge (e.g., Saxenian 1994). The coefficient on commercial isolation remains negative and statistically significant, suggesting a 23.3% decrement in the likelihood of isolated papers being referenced. In column (6) we utilize both city and institution fixed effects, which preserves the negative and statistically significant coefficient on *ISOLATED* and with similar magnitude. As our most conservative specifications, Table 4 suggests that papers isolated with respect to relevant industrial R&D are approximately 10-23% less likely to be referenced.

In Table 5 we consider additional robustness checks and placebo tests. Tables 3 and 4 analyze papers and patents published worldwide, but given our reliance on USPTO data and given the heterogeneity of academic institutions outside of the United States, we restrict our analysis to U.S papers and patents in Table 5, and therefore drop both the *PAPER_US* and *SAME_COUNTRY* covariates. Doing so approximately halves the dataset but allows us to introduce new variables (e.g., *INSTITUTION_TTO*) for which only U.S. data is available. Column (1) corresponds to the same model of that of Table 4 column (4), but for the U.S. only. The fact that a paper was itself patented no longer predicts scientific

references from other patents, and the prestige of the institution does not appear to have a bearing on patterns of scientific referencing. The U.S. corporate patents appear somewhat more likely to reference academic papers in the same city, with a statistical significance at the 10% level. The *ISOLATED* variable is still statistically significant and with magnitude similar to the final column of Table 4.

Table 5 about here

Column (2) and (3) of Table 4 test for two potential mechanisms that might explain in part the impact of isolation. Column (2) investigates whether our main result could be driven by the distribution of academic scientists that are also prolific inventors. Academic scientists that have a large patent stock might be more visible to corporate inventors and their publications might therefore be more likely to be referenced in US patents. We test whether a corresponding author's patent stock predicts whether their paper will be referenced in corporate patents. The *AUTHOR_PATENT_STOCK* variable corresponds to the count of patents awarded by the USPTO to each corresponding author by the calendar year of the simultaneous discovery. As apparent in column (2), this variable does not seem to predict referencing.. Column (3) considers that some academic institutions make more efforts than others to commercialize their research output, and that these efforts might have an impact on corporate inventors' referencing patterns. The Association of University Technology Managers records the year in which U.S.-based universities first set up a technology transfer office, so in column (3) we subset our analysis to U.S.-based universities (i.e., research institutes and other non-university academic institutions are excluded). We introduce the variable *INSTITUTION_TTO* to indicate whether the university had a Technology Transfer Office as of the year of the publication of the simultaneous discovery. The coefficient on the *INSTITUTION_TTO* variable is positive, possibly evidencing that institutional commercialization activity might influence patent reference, but this coefficient is not statistically significant at conventional levels. One reason for this may be sample size, as only 750 patents were at risk of referencing a simultaneous discovery involving at least two US universities; on the other hand, it is possible that the referencing of papers in patents is little affected by the general commercial activity of the paper's source. In any case, the isolation indicator remains statistically significant whether we control for the academic scientist's patent stock or for the existence of a TTO, with magnitude similar to column (1).

In column (4) of Table 5 we test the robustness of our simultaneous-discovery detection algorithm, using the same sample. Above we detailed how we find papers reporting simultaneous discoveries as well as our heuristics for separating out papers that happen to be cited jointly in the academic literature from those that truly represent a simultaneous discovery. That most "twin" papers tend to be published in the same year is somewhat reassuring, but in column (4) of Table 5 we impose a far stricter criterion: that the twin articles be published *back-to-back in the same issue of the same journal*. As noted above, editors often elect to publish multiple papers reporting the same discovery in order to

increase the credibility of the finding. We determine back-to-back publishing by ensuring that two papers are published in the same issue of the same journal, and also that the final page number of one paper (according to Scopus) is within two pages of the starting page number of other paper. Somewhat less than half of our paper twins (41.4%) are published back-to-back, and there are even instances of three papers being published back-to-back-to-back. Column (4) of Table 5 shows that the effect of isolation on the likelihood of being referenced is even stronger for this subsample.

Finally, in column (5) we perform a placebo test. The dyad analyses of Table 3 and in Table 4 are composed of academic papers and non-university patents in order to measure the flow of knowledge from academia to industry. If isolation from relevant industrial R&D affects only these flows and not knowledge diffusion more generally, isolation should not impact flows *within* academia, i.e., to university-owned patents. (Note: we are not measuring here universities patenting the discovery in question—captured by the *PAPER_PATENTED*—variable, but rather references to the academic paper that appear in university-owned patents.) In column (5) we instead create dyads of paper twins and university-assigned patents that cite them, excluding as above self-references by either the authors of the paper or other scientists at the same institution. While the sign of the coefficient on *PAPER_ISOLATED* is negative, it fails to achieve statistical significance at conventional levels ($p<0.247$). As expected considering the way we identified simultaneous discoveries, isolation in our data does not materially impact the flow of academic discoveries to university patents. This placebo test is moreover robust to examining papers and university patents worldwide.

4. Discussion

This paper proposes a methodology to identify frictions in the flow of scientific knowledge between academic research institutions and corporate R&D labs. We focus on simultaneous discoveries—i.e., events in which several teams of scientists share credit for a discovery. Those events constitute a rich setting to study frictions in the flow of knowledge because they expose variance in the exploitation by corporate inventors of the same piece of scientific knowledge produced in different environments. At a time when governments around the world attempt to increase the economic impact of their public investment in science, appetite for new methods to understand this process has increased (see for example scienceofsciencepolicy.net). One contribution of this paper is that it demonstrates how a focus on simultaneous discoveries can be used as a new method to study the circumstances under which inventors exploit new scientific knowledge.

Our analysis of simultaneous discoveries shows that the publication of new scientific knowledge alone does not guarantee that corporate inventors will use it. Frictions can hinder the flow of academic

knowledge to industrial R&D. We focus on the different geographic distributions of institutions of academic research and of industrial R&D. In practice, we measure the flow (or non-flow) of science by observing patent referencing (or not) academic publications disclosing the same discovery but that emerged in different institutions. Besides, since isolation from R&D activity changes across fields and over time, our setting allows us to introduce academic institution-level fixed-effects. In our specification including fixed effects not only for simultaneous discoveries but also cities and institutions, we find that the odds ratio of a patent that refers to one of the paper twins referring to the focal paper is 10-23% lower if it is geographically isolated from the relevant industrial R&D. Interestingly, this effect is not driven by distance alone—therefore indicating that scientific knowledge flows more readily across areas of R&D activity than outside of these areas. Thus, different geographic distributions of institutions of academic research and of industrial R&D appear to create frictions in the dissemination of academic science.

Our findings bear directly on public policy regarding the translation of science. The current distribution of academic research organization might promote equal access to science across geographical areas, but our results suggest that such efforts toward egalitarianism may come at a cost, by systematically complicating firms' exploitation of discoveries made at isolated academic institutions. The variance that we observe in the flow of academic science to industry presents therefore something of a dilemma for science policy. If funding to isolated institutions is in fact less efficient in terms of producing discoveries that are impactful outside of academia, is it rational—and perhaps welfare-enhancing—to purposely channel funding to non-isolated areas? Or should policy-makers introduce measures to specifically promote the dissemination of scientific knowledge produced at isolated institutions so that their results are not neglected by firms?

Isolated institutions may themselves wish to take steps to offset their inherent disadvantage due to location. Given the lower likelihood that relevant firms will take note of their scientists' work, such institutions may want to implement (or enhance) programs to promote their discoveries to industry. Such program might encourage social relationships between academic researchers and industry scientists working in a relevant field. Alternatively, although location may be a given, institutions can set up campuses in less isolated regions, such as Cornell University's "tech" campus on Roosevelt Island in Manhattan or Aalborg University's satellite campus in Copenhagen. Future research may be able to assess the benefits of such steps, which are clearly expensive and potentially disruptive within the institution.

Our findings have implications for the careers of scientists themselves. If, as our results suggest, science of the same nature and quality is less likely to have an impact outside of academia when conducted at an isolated institution, geographic isolation may yield stratification among scientists. Although publications alone might suffice to attain a certain status within the scientific community, researchers at isolated institutions may be denied the popular acclaim and financial rewards that

accompany success in one's research having impact beyond academia. It may be that scientists seeking such rewards already attempt to sort into institutions that are investing in the dissemination of science to industry, but even this might not be enough. Our analysis suggests that an academic institution's ability to have an impact depends critically on the local industrial R&D activity.

Firms may also benefit from exploiting the fact that valuable scientific discoveries emerging from geographically isolated institutions tend to be ignored. This lack of exploitation of academic knowledge might result in missed commercial opportunities, but it could also be rational if access to these discoveries is costly. Our analysis cannot assess whether the existence of those niches constitute an underexploited mine of technological opportunities, but it does raise the question of the possible existence of cost-effective approaches to tap into that knowledge. Whereas many companies may pay attention to discoveries from prestigious universities located near relevant R&D that tend to be written up in the media, firms may gain advantage by paying attention to scientists at isolated universities.

More generally, our analysis points to the importance of continuing research that can further uncover the costs and benefits of the current organization of academic science. The empirical difficulty in studying this topic is considerable because the institutional features of academic science did not emerge in random ways but as a result of complex historical processes (Rosenberg and Nelson 1994; David 2001; Mokyr 2002). The main challenge is therefore one of identification. To address this difficulty, most empirical studies to date have used the combination of large-scale citation analysis with a difference-in-difference approach to causal inference (e.g., Murray and Stern 2007; Agrawal and Goldfarb 2008; Azoulay, Graff Zivin, and Wang 2010; Furman and Stern 2011). Exogenous shocks, however, are not available to answer every important question. This paper attempts to enrich the "empiricist's toolbox" by describing a new approach exploiting the occurrence of simultaneous discoveries in science. This empirical strategy can be used to investigate a number of policy-relevant questions for which no shock is available such as the impact of gender on the direction of inventive activity for instance (Bikard & Fernandez-Matteo 2015). Thus, our hope is that this study will contribute to a better understanding of academic science as an institution by contributing theoretical insights about geographic isolation, but also by establishing the value of simultaneous discoveries as a research setting.

References

Adams, James D. 1990. “Fundamental Stocks of Knowledge and Productivity Growth.” *Journal of Political Economy* 98 (4): 673–702.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2008. “Academic Freedom, Private-Sector Focus, and the Process of Innovation.” *The RAND Journal of Economics* 39 (3): 617–35.

Agrawal, A., and A. Goldfarb. 2008. “Restructuring Research: Communication Costs and the Democratization of University Innovation.” *The American Economic Review* 98 (4): 1578.

Agrawal, Ajay, and Rebecca M. Henderson. 2002. “Putting Patents in Context: Exploring Knowledge Transfer from MIT.” *Management Science* 48 (1).

Alcácer, Juan, and Michelle Gittelman. 2006. “Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations.” *Review of Economics and Statistics* 88 (4): 774–79. doi:i: 10.1162/rest.88.4.774</p>.

Alcácer, Juan, Michelle Gittelman, and Bhaven Sampat. 2009. “Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis.” *Research Policy* 38 (2): 415–27. doi:16/j.respol.2008.12.001.

Audretsch, David B., and Paula E. Stephan. 1996. “Company-Scientist Locational Links: The Case of Biotechnology.” *The American Economic Review* 86 (3): 641–52.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. 2010. “Superstar Extinction.” *The Quarterly Journal of Economics* 125 (2): 549–89.

Babbage, Charles. 1832. *On the Economy of Machinery and Manufactures ... Second Edition Enlarged*. Charles Knight.

Belenzon, Sharon, and Mark Schankerman. 2013. “Spreading the Word: Geography, Policy, and Knowledge Spillovers.” *Review of Economics and Statistics* 95 (3): 884–903. doi:10.1162/REST_a_00334.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2002. “Links and Impacts: The Influence of Public Research on Industrial R&D.” *Management Science* 48 (1): 1–23.

Cozzens, Susan E. 1989. *Social Control and Multiple Discovery in Science: The Opiate Receptor Case*. State University of New York Press.

Dasgupta, Partha, and Paul A. David. 1994. “Toward a New Economics of Science.” *Research Policy* 23 (5): 487–521.

David, Paul A. 2001. “From Keeping ‘Nature’s Secrets’ to the Institutionalization of ‘Open Science.’” Working Paper 01006. Stanford University, Department of Economics. https://ideas.repec.org/p/wop/stanec/01006.html.

Drahl, Carmel. 2014. “Consecutive Journal Publications Illuminate Collaboration And Compromise In Chemistry.” *Chemical & Engineering News* 92 (40). http://cen.acs.org/articles/92/i40/Consecutive-Journal-Publications-Illuminate-Collaboration.html.

Furman, Jeffrey, and Scott Stern. 2011. “Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research.” *American Economic Review* 101 (5): 1933–63.

Galasso, Alberto, and Mark Schankerman. 2014. “Patents and Cumulative Innovation: Causal Evidence from the Courts*.” *The Quarterly Journal of Economics*, November, qju029. doi:10.1093/qje/qju029.

Griliches, Zvi. 1990. “Patent Statistics as Economic Indicators: A Survey.” *Journal of Economic Literature* 28 (4): 1661–1707. doi:10.2307/2727442.

———. 1998. “Introduction to ‘R&D and Productivity: The Econometric Evidence.’” In *R&D and Productivity: The Econometric Evidence*, 1–14. University of Chicago Press. http://www.nber.org/books/gril98-1.

Grossman, Gene M., and Elhanan Helpman. 1993. *Innovation and Growth in the Global Economy*. MIT Press.

Halim, Nadia. 2000. “Bridging Apoptotic Signaling Gaps.” *The Scientist*, August 21. http://www.the-scientist.com/?articles.view/articleNo/12974/title/Bridging-Apoptotic-Signaling-Gaps/.

Harris, Gardiner. 2011. “Federal Research Center Will Help Develop Medicines.” *The New York Times*, January 22, sec. Health / Money & Policy. http://www.nytimes.com/2011/01/23/health/policy/23drug.html.

Hellmann, Thomas. 2007. “The Role of Patents for Bridging the Science to Market Gap.” *Journal of Economic Behavior & Organization* 63 (4): 624–47. doi:10.1016/j.jebo.2006.05.013.

Henderson, Rebecca M., Adam B. Jaffe, and Manuel Trajtenberg. 1998. “Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988.” *Review of Economics and Statistics* 80 (1): 119–27. doi:10.1162/003465398557221.

———. 2005. “Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment.” *The American Economic Review* 95 (1): 461–64.

Jaffe, Adam B. 1989. “Real Effects of Academic Research.” *The American Economic Review* 79 (5): 957–70.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca M. Henderson. 1993. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” *The Quarterly Journal of Economics* 108 (3): 577–98. doi:10.2307/2118401.

Jensen, Richard, and Marie Thursby. 2001. “Proofs and Prototypes for Sale: The Licensing of University Inventions.” *The American Economic Review* 91 (1): 240–59.

Lampe, Ryan. 2012. “Strategic Citation.” *Review of Economics and Statistics* 94 (1): 320–33. doi:10.1162/REST_a_00159.

Lemley, Mark A. 2007. “Should Patent Infringement Require Proof of Copying?” *Michigan Law Review* 105 (7): 1525–36.

Li, Guan-Cheng, Ronald Lai, Alexander D’Amour, David M. Doolin, Ye Sun, Vetle I. Torvik, Amy Z. Yu, and Lee Fleming. 2014. “Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975–2010).” *Research Policy* 43 (6): 941–55. doi:10.1016/j.respol.2014.01.012.

MacGarvie, Megan. 2006. “Do Firms Learn from International Trade?” *Review of Economics and Statistics* 88 (1): 46–60. doi:10.1162/rest.2006.88.1.46.

Mansfield, Edwin. 1998. “Academic Research and Industrial Innovation: An Update of Empirical Findings.” *Research Policy* 26 (7-8): 773–76. doi:10.1016/S0048-7333(97)00043-7.

Marshall, A. 1895. *Principles of Economics*. Macmillan.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press.

Murray, Fiona. 2002. “Innovation as Co-Evolution of Scientific and Technological Networks: Exploring Tissue Engineering.” *Research Policy* 31 (8-9): 1389–1403.

Murray, Fiona, and Scott Stern. 2007. “Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge?: An Empirical Test of the Anti-Commons Hypothesis.” *Journal of Economic Behavior & Organization* 63 (4): 648–87. doi:10.1016/j.jebo.2006.05.017.

Nelson, Richard R. 1959. “The Simple Economics of Basic Scientific Research.” *Journal of Political Economy* 67 (3): 297–306.

Nelson, Richard R. 1982. “The Role of Knowledge in R&D Efficiency.” *The Quarterly Journal of Economics*, The Quarterly Journal of Economics, 97 (3): 453–70.

Niehans, Jurg. 1995. “Multiple Discoveries in Economic Theory.” *European Journal of the History of Economic Thought* 2 (1): 1. doi:Article.

Ogburn, William F., and Dorothy Thomas. 1922. “Are Inventions Inevitable? A Note on Social Evolution.” *Political Science Quarterly* 37 (1): 83–98.

Roach, Michael, and Wesley M. Cohen. 2013. “Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research.” *Management Science* 59 (2): 504–25. doi:10.1287/mnsc.1120.1644.

Romer, Paul M. 1990. “Endogenous Technological Change.” *Journal of Political Economy* 98 (5 pt 2). http://www.dklevine.com/archive/refs42135.pdf.

Rosenberg, Nathan. 1994. *Exploring the Black Box: Technology, Economics, and History*. Cambridge University Press.

Rosenberg, Nathan, and Richard R. Nelson. 1994. “American Universities and Technical Advance in Industry.” *Research Policy* 23 (3): 323–48. doi:10.1016/0048-7333(94)90042-6.

Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press.

Singh, Jasjit, and Ajay Agrawal. 2011. “Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires.” *Management Science* 57 (1): 129–50. doi:10.1287/mnsc.1100.1253.

Singh, Jasjit, and Matt Marx. 2013. “Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity.” *Management Science* 59 (9): 2056–78. doi:10.1287/mnsc.1120.1700.

Sohn, Eunhee. 2014. “The Endogeneity of Academic Science to Local Industrial R&D.” *Academy of Management Proceedings* 2014 (1): 11413. doi:10.5465/AMBPP.2014.286.

Stephan, Paula E. 1996. “The Economics of Science.” *Journal of Economic Literature* 34 (3): 1199–1235.

Thompson, Peter, and Melanie Fox-Kean. 2005. “Patent Citations and the Geography of Knowledge Spillovers: A Reassessment.” *The American Economic Review* 95 (1): 450–60.

Vermont, Samson. 2006. “Independent Invention as a Defense to Patent Infringement.” *Michigan Law Review* 105 (3): 475–504.

Weintraub, E. Roy. 2011. “Retrospectives: Lionel W. McKenzie and the Proof of the Existence of a Competitive Equilibrium.” *Journal of Economic Perspectives* 25 (2): 199–215. doi:10.1257/jep.25.2.199.

Zucker, Lynne G., Michael R. Darby, and Marilynn B. Brewer. 1998. “Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises.” *The American Economic Review* 88 (1): 290–306.

Table 1: Variable definitions.

paper-level variables	
NUMREFS	Number of patents referencing the focal paper as non-patent prior art.
ISOLATED	The academic institution in which the paper's corresponding author is based is not within 50 miles (commuting distance) of any city with at least five patents and more than 2% of patenting in relevant technological subclasses.
PAPER_US	Paper's corresponding address is in U.S.
PAPER_PATENTED	One of paper's authors was granted a patent on the discovery reported by the paper. (Note: not counted in NUMREFS or PATREF.)
PAPER_ JIF	Impact factor of journal paper was published in, calculated as five-year running average as of 2009 (logged).
INSTITUTION_PRESTIGE	Number of papers published in top 15 scientific journals with corresponding address at the same institution as the focal paper. (logged)
INSTITUTION_TTO	The paper's institution had established a technology-transfer office as of the publication year of the paper (U.S. institutions only.)
paper-patent dyad	
REFERENCED	Focal paper is referenced by patent in the paper-patent dyad.
TIME_LAG	Time lag between publication of focal paper and possibly-referencing patent.
SAME_CITY	Paper and patent are in the same city.
SAME_MSA	Paper and patent are in same Metropolitan Statistical Area (using 2003?? CBSA definitions).
SAME_STATE	Paper and patent are in the same state (North America only.)
SAME_COUNTRY	Paper and patent are in the same country.
DIST	Spatial distance (in miles) between paper and patent, adjusted for curvature of the earth. (L) Some models instead present dummies for various categories of separation: 0-10 miles, 11-20, etc.

Table 2: Descriptive statistics and correlations for simultaneous discoveries**.**

	Variable	Obs	Mean	Stdev	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	*NUMREFS*	2,889	32.216	26.014	1.000	102	1.000												
(2)	*ISOLATED*	2,889	0.724	0.447	0.000	1	-0.296	1.000											
(3)	*PAPER_US*	2,889	0.683	0.465	0.000	1	0.173	-0.384	1.000										
(4)	*PAPER_PATENTED*	2,889	0.299	0.458	0.000	1	0.130	-0.109	0.071	1.000									
(5)	*PAPER_JIF*	2,889	3.132	0.637	0.000	3.959	0.054	-0.020	-0.131	0.011	1.000								
(6)	*INSTITUTION_PRESTIGE*	2,889	4.199	1.628	0.693	6.551	0.010	-0.253	0.448	0.065	0.000	1.000							
(7)	*INSTITUTION_TTO*	959	0.667	0.471	0.000	1	-0.112	0.176	0.046	-0.090	-0.128	0.482	1.000						
(8)	*REFERENCED*	2,889	0.633	0.482	0.000	1	0.009	-0.144	0.115	0.142	0.021	0.012	-0.114	1.000					
(9)	*TIME_LAG*	2,889	4.886	3.291	0.000	17	0.043	-0.122	-0.006	-0.076	0.102	0.143	0.097	0.011	1.000				
(10)	*SAME_CITY*	2,889	0.036	0.185	0.000	1	0.079	-0.241	0.127	0.196	0.093	-0.009	-0.155	0.112	-0.060	1.000			
(11)	*SAME_MSA*	1,503	0.131	0.338	0.000	1	0.031	-0.231	N/A	0.101	0.202	0.060	-0.119	0.148	-0.018	0.695	1.000		
(12)	*SAME_STATE*	1,649	0.158	0.365	0.000	1	-0.019	-0.308	0.110	0.034	0.173	0.092	-0.151	0.125	-0.021	0.588	0.806	1.000	
(13)	*SAME_COUNTRY*	2,889	0.530	0.499	0.000	1	0.165	-0.234	0.681	0.092	-0.076	0.300	-0.052	0.090	-0.044	0.181	N/A	0.135	1.000
(14)	*DIST*	2,889	7.197	1.957	0.000	9.264	-0.111	0.243	-0.298	-0.139	-0.097	-0.156	0.139	-0.106	0.031	-0.696	-0.870	-0.790	-0.504

Notes: Variables are defined in Table 1. Observations are constructed for all combinations of all twin academic papers and potentially-referencing corporate patents for all simultaneous discoveries where at least one of the twin papers is referenced by some patent. *SAME_MSA* is defined only in the U.S. and thus does not vary according to *PAPER_US* or *SAME_COUNTRY*. *SAME_STATE* however is defined for Canadian provinces and thus is not missing.

Table 3: Isolation and referencing of "twin" simultaneous-discovery articles by corporate patents.

	(1)	(2)	(3)	(4)
Dep Var =	*NUMREFS*	*REFERENCED*	*REFERENCED*	*REFERENCED*
ISOLATED	-9.802***	-0.174**	-0.160**	-0.148**
	(1.581)	(0.0701)	(0.0673)	(0.0626)
PAPER_US	-1.442	0.116*	0.100*	0.113*
	(1.199)	(0.0594)	(0.0572)	(0.0633)
PAPER_PATENTED	4.268***	0.126**	0.119**	0.120**
	(1.250)	(0.0505)	(0.0486)	(0.0476)
PAPER_JIF	-6.658	0.121	0.121	0.107
	(4.417)	(0.176)	(0.180)	(0.174)
PAPER_JIF^2	1.609*	-0.0178	-0.0186	-0.0160
	(0.832)	(0.0331)	(0.0339)	(0.0326)
INSTITUTION_PRESTIGE	-1.099	-0.106	-0.115*	-0.105*
	(1.437)	(0.0692)	(0.0643)	(0.0573)
INSTITUTION_PRESTIGE^2	0.158	0.0106	0.0117	0.0105
	(0.187)	(0.0100)	(0.00916)	(0.00811)
TIME_LAG		0.00230	0.00286	0.00315
		(0.00422)	(0.00428)	(0.00435)
DIST			-0.0183	
			(0.0112)	
SAME_CITY				0.0973
				(0.111)
SAME_COUNTRY				0.00311
				(0.0484)
DIST1-10				0.260***
				(0.0861)
DIST11-50				-0.0213
				(0.0997)
DIST50-200				-0.0969
				(0.139)
DIST201-1000				-0.0679
				(0.0832)
DIST1001-6000				-0.109
				(0.0752)
CONSTANT	23.17***	0.675***	0.830***	0.753***
	(6.309)	(0.253)	(0.272)	(0.257)
simultaneous discovery fixed effects	no	yes	yes	yes
# observations	380	2889	2889	2889
unit of analysis	paper	paper-patent dyad	paper-patent dyad	paper-patent dyad
# of simultaneous discoveries	187	187	187	187

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Notes: Dependent variable is either the number references the focal paper receives from corporate patents (col. 1) or whether a paper-patent dyad is linked by an actual reference (cols. 2-4). *ISOLATED* refers to a paper at an institution not within 50 miles of a "hub" of R&D activity for the patent subclasses associated with its simultaneous discovery. Controls are defined in Table 1. Standard errors are clustered at the level of the simultaneous discovery. All models use papers and patents published worldwide.

Table 4: Isolation with fixed effects for the referencing institution and/or its city.

	(1)	(2)	(3)
Dep Var =	*REFERENCED*	*REFERENCED*	*REFERENCED*
ISOLATED	-0.0968**	-0.233***	-0.205***
	(0.0460)	(0.0452)	(0.0553)
PAPER_US	0.143*	0.0882	
	(0.0754)	(0.115)	
PAPER_PATENTED	0.154***	0.174***	0.142***
	(0.0365)	(0.0324)	(0.0391)
PAPER_JIF	0.0815	-0.157	0.168
	(0.284)	(0.246)	(0.311)
PAPER_JIF^2	-0.00543	0.0342	-0.0220
	(0.0513)	(0.0440)	(0.0556)
INSTITUTION_PRESTIGE		-0.182***	
		(0.0469)	
INSTITUTION_PRESTIGE^2		0.0237***	
		(0.00589)	
TIME_LAG	0.00579*	0.00717**	0.00627*
	(0.00323)	(0.00319)	(0.00328)
SAME_CITY	0.142*	0.124	0.153*
	(0.0758)	(0.0763)	(0.0882)
SAME_COUNTRY	-0.0399	-0.0117	-0.0235
	(0.0278)	(0.0278)	(0.0282)
DIST1-10	0.249***	0.201**	0.237**
	(0.0939)	(0.0956)	(0.0961)
DIST11-50	0.129*	0.0463	0.125*
	(0.0675)	(0.0670)	(0.0687)
DIST50-200	0.0302	0.00163	0.0199
	(0.0680)	(0.0675)	(0.0704)
DIST201-1000	-0.0117	-0.0209	0.00660
	(0.0522)	(0.0511)	(0.0542)
DIST1001-6000	-0.00185	-0.0225	0.00768
	(0.0451)	(0.0447)	(0.0472)
CONSTANT	-0.0989	0.596	0.445
	(0.499)	(0.587)	(0.685)
institution FE	yes	no	yes
city FE	no	yes	yes
# observations	2300	2534	2225
# simultaneous discoveries	168	182	161

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Notes: Dependent variable reports whether a paper-patent dyad is linked by an actual reference. *ISOLATED* refers to a paper published by an institution not within 50 miles of a "hub" of R&D activity for the patent subclasses associated with its simultaneous discovery. Controls are defined in Table 1. Standard errors are clustered at the level of the simultaneous discovery. All models use worldwide papers and patents.

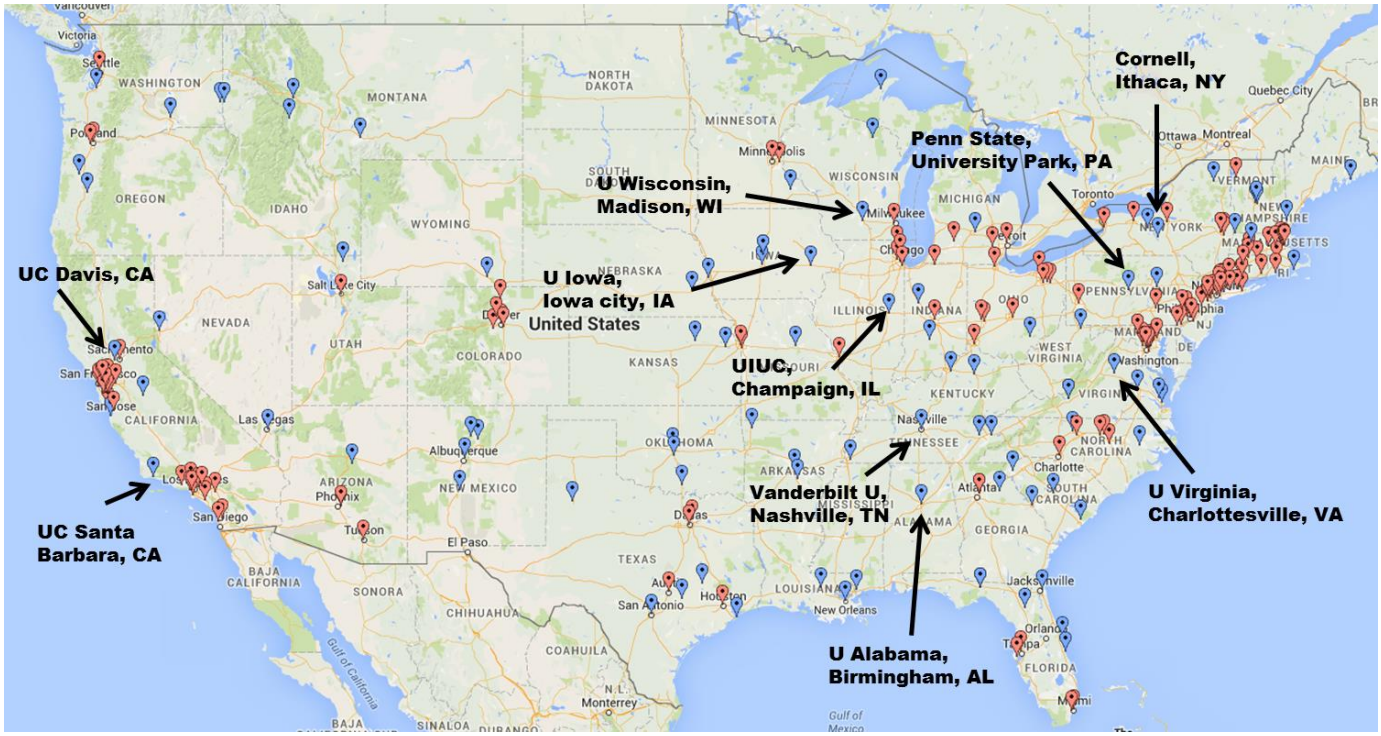
Table 5: Robustness and placebo tests for isolation.

	(1)	(2)	(3)	(4)	(5)
Dep Var =	*REFERENCED*	*REFERENCED*	*REFERENCED*	*REFERENCED*	*REFERENCED*
ISOLATED	-0.257***	-0.265***	-0.283**	-0.623***	-0.0982
	(0.0821)	(0.0786)	(0.125)	(0.206)	(0.0848)
PAPER_PATENTED	0.0951	0.0986	0.0773	0.193	-0.00663
	(0.0685)	(0.0699)	(0.0828)	(0.121)	(0.0611)
PAPER_JIF	-0.167	-0.191	-0.819	0.654	-1.567***
	(0.403)	(0.403)	(0.792)	(0.713)	(0.513)
PAPER_JIF^2	0.0358	0.0396	0.155	-0.141	0.295***
	(0.0736)	(0.0739)	(0.145)	(0.122)	(0.0895)
INSTITUTION_PRESTIGE	-0.0114	-0.0110	0.168	0.119	-0.00692
	(0.0773)	(0.0775)	(0.172)	(0.189)	(0.0715)
INSTITUTION_PRESTIGE^2	-0.00388	-0.00396	-0.0216	-0.0143	-0.00324
	(0.0101)	(0.0101)	(0.0218)	(0.0256)	(0.00985)
AUTHOR_PATENT_STOCK		-0.00189			
		(0.00611)			
INSTITUTION_TTO			0.0206		
			(0.127)		
TIME_LAG	0.0104*	0.0102*	0.0123	-0.00908	0.00479
	(0.00597)	(0.00580)	(0.00911)	(0.00716)	(0.00441)
SAME_CITY	0.213*	0.221*	0.263*	0.0422	0.331***
	(0.118)	(0.126)	(0.149)	(0.0327)	(0.102)
SAME_MSA	0.170	0.171	0.313	-0.210	-0.209
	(0.202)	(0.203)	(0.283)	(0.217)	(0.173)
SAME_STATE	-0.000111	0.00206	0.0130	0.291*	0.0885
	(0.108)	(0.109)	(0.0940)	(0.161)	(0.125)
DIST1-10	0.278	0.289	0.334*	-0.105	0.158
	(0.195)	(0.200)	(0.171)	(0.158)	(0.183)
DIST11-50	0.163	0.176	0.227	-0.148	0.0691
	(0.192)	(0.199)	(0.168)	(0.158)	(0.123)
DIST50-200	0.203	0.220	0.490	-0.155	0.0715
	(0.295)	(0.297)	(0.342)	(0.400)	(0.231)
DIST201-1000	0.332	0.348	0.558	-0.219	-0.0111
	(0.291)	(0.295)	(0.342)	(0.279)	(0.231)
DIST1001-6000	0.237	0.255	0.265	-0.0346	0.0602
	(0.288)	(0.294)	(0.324)	(0.279)	(0.237)
CONSTANT	1.148*	0.840	2.330**	0.410	2.686***
	(0.594)	(0.605)	(1.136)	(1.136)	(0.775)
paper twin qualifier	N/A	N/A	universities	back-to-back	N/A
patents assigned to	firms	firms	firms	firms	universities
# observations	1503	1503	750	483	1099
# simultaneous discoveries	137	137	69	62	200

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Notes: All models use U.S.-based papers and patents. Standard errors are clustered at the level of the simultaneous discovery. Variables are defined in Table 1.

Figure 1. Geographic Isolation of Top Academic Research Institutions in the US



Notes: Blue pins are geographically isolated from inventors (Fewer than 5000 patents were awarded to inventors living within a 25-mile radius). The ten institutions that are labeled are the most productive isolated institutions as measured by the number of papers that they published in top-15 impact factors journals between 2000 and 2010.

Figure 2: Construction of simultaneous-discovery dataset of dyads between each paper reporting the simultaneous discovery and each patent that referenced any of the papers reporting the simultaneous discovery. Each line represents a dyad in the dataset. Solid lines represent actual references while dotted lines represent unrealized references.