I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13419

A Firm Scientific Community: Industry Participation and Knowledge Diffusion

Stefano Baruffaldi
Felix Poege

JUNE 2020

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13419

A Firm Scientific Community: Industry Participation and Knowledge Diffusion

Stefano Baruffaldi
University of Bath and MPI for Innovation and Competition

Felix Poege
MPI for Innovation and Competition and IZA

JUNE 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A Firm Scientific Community: Industry Participation and Knowledge Diffusion[*]

We study the diffusion of knowledge from scientists to firms within scientific communities. We build on a unique dataset on conference proceedings as "paper trail" of almost all relevant conference series in computer science since 1996. More than 5000 firms appear as conference sponsors or as affiliations in proceedings. Their participation is concentrated in the highly ranked conferences and their scientific contributions are on average highly cited. We exploit direct flights as an instrumental variable for the participation choice of scientists between a conference where a firm participates and other similar conferences. The participation in the same conference has a positive causal effect on knowledge diffusion to the firm's scientific and inventive activities. Additional analyses suggest that interactions and collaborations with scientists that remain external to the firm are likely a key mechanism of this diffusion. The effects are remarkably stronger the larger the firm's investments in participation.

JEL Classification: O33, O32, D22

Keywords: knowledge diffusion, corporate science, scientists, conferences

Corresponding author:
Felix Poege
Max Planck Institute for Innovation and Competition
Marstallplatz 1
80539 Munich
Germany
E-mail: Felix.Poege@ip.mpg.de

[*] We are grateful for helpful comments to Christian Fons-Rosen, Fabian Gaessler, Dietmar Harhoff, Rossella Salandra, Ammon Salter, Markus Simeth, Florenta Teodoridis, Anne Ter Wal, Stefan Wagner and Martin Watzinger. We want to thank the participants of the REER, AOM, DRUID, EPIP, SMYE and Innopat conferences and seminars at LMU and TU Munich, CREI-Bath, SKEMA, University of Bath and Imperial College London for their comments. Excellent research assistance was provided by Gregor Bös, Bernhard Kassner, Halime Kobra Kocak, Natalia Shvets and Kathrin Wernsdorf. Stefano H. Baruffaldi acknowledges financial support from the Swiss Science National Foundation (Reference No.: P2ELP1-161847). Felix Poege acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG) (Reference No.: GA 2511/1-1).

1 Introduction

The Advances in Neural Information Processing Systems (NeurIPS) conference is a leading academic scientific conference in Machine Learning (ML). In 2017, Google featured as an official sponsor and, with 75 published proceedings, was, by far, the most represented affiliation of scientists at the event. Other companies such as Microsoft, IBM or Tencent follow not too distant down in the list.[1] The diffusion of knowledge from science to industry has been a topic of enduring interest for economists due to its importance for firms' productivity and economic growth (Aghion, Dewatripont, and Stein, 2008; Arora, Belenzon, and Sheer, 2020; Arora and Gambardella, 1994; Grossman and Helpman, 1993; Jaffe, 1989; Romer, 1990). Knowledge diffusion tends to be localized in the geographical and technological space (Audretsch and Feldman, 1996; Belenzon and Schankerman, 2013; Jaffe, Trajtenberg, and Henderson, 1993) and firms have to make specific investments to increase their capacity to absorb external knowledge (Cassiman and Veugelers, 2002; Cohen and Levinthal, 1989). Firms' connections and interactions within scientific communities can be an important channel of diffusion of scientific knowledge and some authors led the way to this line of investigation (Audretsch and Stephan, 1996; Cockburn and Henderson, 1998; Hicks, 1995; Rosenberg, 1990). However, the diffusion of knowledge from science to firms participating in scientific communities remains largely unexplored.

Social scientists have long been aware of the role of scientific communities for the production and diffusion of knowledge (Cetina, 1999; Crane, 1974; Dasgupta and David, 1994; Merton, 1973; Stephan, 1996). Modern science is shaped by groups of scientists that organize around common interests. These communities gather regularly at conferences to share information and findings. For individuals and organizations, active participation to a community is also a way to gain reputation and comply with social norms of mutual reciprocity that ease professional and social relationships (Charness, Rigotti, and Rustichini, 2007; Haeussler, 2011; Haeussler et al., 2014). Therefore, scientists within the same scientific communities engage in a process of socialization and establishment of personal connections that channel knowledge diffusion (Boudreau et al., 2017; Campos, Lopez de Leon, and McQuillin, 2018; Chai and Freeman, 2019; Lopez de Leon and McQuillin, 2020). The conjecture that similar mechanisms are at play when firms engage with scientific communities is proximate. Specific evidence has been rare because it is difficult to assemble relevant data on a large scale and to empirically isolate the effect of the participation in the same scientific community from the factors that, in the first place, determine its choice.[2]

In this paper, we ask whether knowledge diffuses more likely from scientists to firms that par-

[1] Further information available at https://nips.cc/Conferences/2017. In recent years, intense participation of industry to is common to most scientific conferences in Machine Learning (ML). This trend is debated also by internal observers to the scientific community (see for instance: http://www.argmin.net/2018/08/09/co-employment/)

[2] There are notable exceptions: Gittelman and Kogut (2003) and Gittelman (2007) look at the trade off between scientific and innovation performance of firms in biotechnology; Cassiman and Veugelers (2002, 2006) and Simeth and Raffo (2013) study how the relevance of academic knowledge influence firms' strategy; Vlasov, Bahlmann, and Knoben (2016) look at the association between small and medium-size firms' innovation and the knowledge diversity of conferences where they participate.

ticipate in the same scientific communities. We assemble a unique dataset of all relevant scientific conferences in Computer Science (CS), worldwide, from 1996 to 2015. CS is an ideal setting for our study, additional to its economic relevance (Brynjolfsson and Hitt, 2003; Nelson, 1962). Detailed information on conference series is captured in well-curated datasets, due to the importance of conference proceedings for scientists (Franceschet, 2010).[3] We leverage information from conference proceedings as a "paper trail" of the participation of scientists and firms to distinct scientific communities.

We are mainly interested in and able to observe participation defined as the active contribution to scientific conferences, as opposed to passive attendance. We capture the two possible forms of firms' participation: the authorship of proceedings by firms' affiliated scientists and the sponsorship of conferences. Firms' scientists as authors of proceedings participate in a similar way as academic scientists do. Sponsorship entails a specific financial contribution to the conference and gives in turn additional opportunities to promote the reputation of the firm, space to expose its research and for hiring activities, and additional entrance tickets. Our understanding of this phenomenon is enriched by interviews with participants at two important conferences.[4] We discuss in the paper how both authorship and sponsorship closely reflect a significant investment and engagement. A higher number of proceedings and sponsorship decisions serve also as a proxy of the intensity of participation of the firm, implying larger investments and stronger presence at a conference.

We estimate the effect on knowledge diffusion to firms of the participation of external scientists to the same conferences where firms participate. In our main specifications, we capture knowledge diffusion using citations from firms' publications or patents to scientists' proceedings. For a given sample of proceedings at a conference where a firm participates, we consider a counterfactual group of proceedings from another comparable conference (in the same year, the same field, comparable quality and size). Participation is then defined as the participation of other scientists external to the focal firm, revealed by the actual presence of a proceeding in the same conference where the firm participates. To establish causality, our econometric models isolate exogenous variation in the probability of this participation arising from the availability of direct flights, as proxy of general costs of transportation (Catalini, Fons-Rosen, and Gaulé, 2020; Giroud, 2013). Different levels of fixed effect controls (FE) rule out variation that may correlate with both the availability of direct flights and the dependent variables (Giroud, 2013): quality of conference series, pair-level characteristics of firms and the location of scientists' affiliations; time-specific idiosyncratic shocks to scientists' locations and firms.

Our data portray 7298 conferences from 1042 conference series and 5470 participating firms. Firms' authorship occurs in 88.3% of the conferences and sponsorship in 26.6%, at a rather constant

[3] As we further discuss in the paper, in CS, conference proceedings are valuable publications for scientists. Multiple presentations of the same paper at multiple conferences is rare and difficult. "No shows" - the unjustified lack of at least one presenter - is considered unethical and may imply the elimination of the paper from the conference proceedings.

[4] We visited the ECCV conference 2018 in Munich and the NeurIPS conference 2019 in Vancouver. We discuss the insights from interviews with firm scientists and HR representatives in the paper and, more extensively, in the dedicated appendix D.

rate over time. Firm participation is concentrated in the conference series of the highest quality. Firm authored proceedings are a relatively small share of the total (10.9%) but are, on average, highly cited. The participation frequency and intensity is highly skewed, with the top firms being responsible for the majority of the observed participation.

The empirical analysis provides evidence of a much higher probability of knowledge diffusion from scientists to firms after participation in the same conferences. The main estimates are based on a sample of more than 5 million observations at firm and conference proceedings pair-level. While large in absolute numbers, actual citations are rare in this sample. The probability of scientific citations from firms, after a conference, increases by 1.3 to 2.1 percentage points (p.p.) in our preferred specifications. This is large considering as a benchmark, for instance, that the familiarity of the firm with the authors' previous work, measured by previous citations, is associated with a 2.9 p.p increase. The increase in science citations to past proceedings of scientists is also significant and similar in magnitude. The increase in the probability of patent citations to conference proceedings at the conference is small and not statistically significant. On the contrary, the probability of patent citations to previous publications of the authors of the conference proceedings increases significantly, by 11 p.p, after the conference. The presence of previous patent citations by the firm to the same proceedings correlates with this dependent variable by 19.5 p.p.

We discuss and explore in additional analyses the mechanisms of diffusion. First, we expect actual interactions with scientists to play an important role (Audretsch and Stephan, 1996; Cockburn and Henderson, 1998). The relatively large effect on citations to past publications of scientists, particularly for patent citations, is already supportive of this hypothesis. To provide direct evidence, we study the effect of participation on the probability of future scientific collaborations between the firm and the scientists at conferences, captured by future coauthored publications. We find a strongly positive effect on this variable. Similarly, we study the effect of participation on the probability of the mobility of scientists, as measured by changes in scientists' affiliations in publications. Scientists hired by the firm at the conference can transfer knowledge directly, and can also be responsible for the observed citations (Almeida and Kogut, 1999; Stephan, 1996). Interestingly, on average, we find a rather null effect on hiring.

Second, knowledge sharing within scientific communities, in the absence of market mechanisms, is governed by a system based on reputation and mutual reciprocity (Dasgupta and David, 1994; Haeussler et al., 2014; Stephan, 1996). Accordingly, firms that make it to the top of the prestige distribution would obtain the highest returns. To explore this hypothesis, we study the heterogeneity of the effects by the intensity of the participation of the firm. Stronger contributions would increase reputation, and further ease knowledge sharing and the establishment of collaborations. Conversely, low or intermediary levels of participation may suffice if only the direct access to the information presented at conferences matters. We find that all effects increase greatly with the intensity of participation of the firm. The effects are weak or zero for low levels of participation, where a firm has only one proceeding presented and is not a conference sponsor. Sponsorship without authorship is also associated with weaker effects. At the opposite extreme, firms that both sponsor the conference and author several proceedings, participation has the strongest effects, arriving to be positive and significant also on patent citations to the same conference proceedings presented at the conference and on hiring. We find a similar pattern using an indicator of firm size of research investments, based on the number of the firm's active scientists.

The paper outlines a series of robustness analyses on the main results. We use text similarity as an alternative to scientific citations. This addresses the concern that citations may emerge merely due to the saliency of some references, without an actual shift of the content of research. Yet, we also find a strong and significant effect based on text similarity measures. To test the validity of our models we provide evidence that the predetermined level of citations and text similarity, before the conference, are uncorrelated with our variable of interests within our model specification. Specifically, within our models, participation shows no significant correlation with scientific citations, patent citations and research similarity based on documents of firms and scientists published before the conference. We also test the robustness of the models to the inclusion of several additional FE controls. Finally, we verify that alternative clustering of standard errors has no implications for the significance of the results.

The paper builds on and contributes to two literature streams. First, our results relate to the literature on corporate science (Arora, Belenzon, and Sheer, 2020; Cockburn and Henderson, 1998; Gittelman, 2007; Hicks, 1995; Rosenberg, 1990) and the external knowledge sources of firm innovation (Cassiman and Veugelers, 2002, 2006; Simeth and Raffo, 2013). Corporate science has paved the way to breakthroughs in the past, but recent studies have documented a decline in firms' investments in science (Arora, Belenzon, and Patacconi, 2018). At the same time, scientific knowledge firmly remains a driver of technology and firm value (Arora, Belenzon, and Patacconi, 2018; Fleming, Greene, et al., 2019; Simeth and Cincera, 2016). The participation to scientific communities is endemic to modern science and, in survey data, conferences score highly in importance as channels of knowledge diffusion from public research to industry R&D (Cohen, Nelson, and Walsh, 2002). Our results offer novel evidence to understand the relevance of participation to scientific communities and the rationale for corporate science, more broadly. Investments in active participation can be compatible with firms' objectives because connections with scientific communities help accessing external scientific knowledge. Participation to a conference is a relatively accessible investment. However, not only is internal research likely complementary (Cassiman and Veugelers, 2002, 2006), but participation can be subject steep increasing returns to investments.

The paper also contributes to the literature on knowledge diffusion (Audretsch and Stephan, 1996; Eeckhout and Jovanovic, 2002; Jaffe, 1989; Stein, 2008). Extant work has demonstrated the localized nature of knowledge spillovers showing that also opportunities for temporary proximity can have large effects (Boudreau et al., 2017; Campos, Lopez de Leon, and McQuillin, 2018; Chai and Freeman, 2019; Lopez de Leon and McQuillin, 2020). We bring attention to the role of organizations' investments as antecedents to such opportunities. Specifically, firms put in place non-market strategies to acquire external knowledge, based on investments to position themselves within external scientific communities. The findings reinforce the evidence that temporary proximity can be a vehicle of knowledge diffusion. However, they defy the notion of knowledge diffusion as free and evenly distributed spillovers (Breschi and Lissoni, 2001; Eeckhout and Jovanovic, 2002).

The evidence is compatible with an important role of reputation and prestige of organizations and the centrality of social relationships within communities, that go beyond the effect of proximity and search costs, and yield increasing returns on participation investments. An uneven pattern of diffusion emerges, that more plausibly leads to divergence and concentration, rather than convergence, in firms innovative productivity (Andrews, Criscuolo, and Gal, 2015).

2 Participation and knowledge diffusion in scientific communities

A fundamental tenant in the economics of innovation is that knowledge does not flow spontaneously (Akcigit et al., 2018; Jaffe, Trajtenberg, and Henderson, 1993). Geographical and technological boundaries, combined with an ever-increasing body of cumulated knowledge, determine high search costs for individuals and firms (Jones, 2009; Rosenkopf and Nerkar, 2001). Science operates in communities, united by a common interest, that share knowledge systematically (Cetina, 1999; Crane, 1974). As such, the structure of science constitutes a way to organize and navigate existing knowledge (Fleming and Sorenson, 2004). This structure is manifested in international conferences, that turn into loci of controlled spillovers of specialized, up to date, and temporary localized knowledge (Bathelt and Cohendet, 2014). Participation helps abating search costs within a body of knowledge which is defined along some dimension (e.g. scientific area), but otherwise scattered in a variety of dispersed, and possibly unknown, locations (Maskell, 2014). Recent literature contributions have found empirical evidence on the effect of temporary proximity at conferences on the probability of knowledge flows and collaborations between scientists (Boudreau et al., 2017; Campos, Lopez de Leon, and McQuillin, 2018; Chai and Freeman, 2019; Lopez de Leon and McQuillin, 2020).

Active participation to the same community is helpful, because the mere physical presence or reading of proceedings at conferences may not naturally translate into knowledge diffusion, which is instead embodied in effective voluntary interactions (Hicks, 1995). Knowledge production within a community is a collective process where active participation and compliance to social norms are expected (Merton, 1973). Reciprocity and the reputation gained from previous contributions become the currency that aligns incentives for knowledge production and sharing (Dasgupta and David, 1994; Stephan, 1996). The willingness to share knowledge may remain anchored to cost and benefits considerations based on expected reciprocity (Bobtcheff, Bolte, and Mariotti, 2017; Mukherjee and Stern, 2009; Stein, 2008). At the same time, sharing decisions are not based exclusively on contingent negotiations but rely on a sense of community and the prestige of individuals and organizations. There is experimental evidence that group identity (Charness, Rigotti, and Rustichini, 2007; Chen and Li, 2009) and status clues (Bhattacharya and Dugar, 2014) influence cooperative behavior of individuals. Survey evidence also shows that both academic and industry scientists take into account expected reciprocity and the perceived conformity to the norms of science in knowledge sharing decisions (Haeussler, 2011; Haeussler et al., 2014).

This theoretical perspective has implications for the expected role of participation of firms to scientific communities and the understanding of the related mechanisms of diffusion. Firms need to make investments in absorptive capacity to take advantage of external knowledge (Cohen and Levinthal, 1989). As science unfolds as a sequence of inter-independent hypotheses on cause and effect relationships, actively engaging in science is instrumental in making sense of intermediary and external results (Arora and Gambardella, 1994; Hellmann and Perotti, 2011). However, the role of absorptive capacity alone justifies internal research. If the existence of search costs was the only reason, participation in external scientific communities would be subject to evident decreasing returns to investments. The trade-off between positive and negative spillovers due to disclosure would likely remain binding (Cassiman and Veugelers, 2002, 2006; Laursen and Salter, 2006) and excessive focus on science may divert the firm from technological output and profit objectives (Gittelman, 2007; Gittelman and Kogut, 2003).

Conversely, firms may face steep increasing returns to investments in participation on knowledge diffusion. As a consequence of the accumulation of prestige as a form of capital (Matthew effect), social structures of scientific communities are highly skewed, with few individuals and institutions in positions of great influence (Merton, 1973; Stephan, 1996). Scientific meetings become not simply temporary gatherings for knowledge sharing, but also regular occasions to demonstrate participation and commitment and, ultimately, the confrontation fields where these social structures gradually take shape. Individuals and institutions at the far end of the prestige distribution obtain disproportional gains in visibility, and are in a privileged position to collaborate and absorb knowledge (Gittelman, 2007; Gittelman and Kogut, 2003). Competition for these positions, and the notion that existing social structures may require substantial effort to be scratched, justify intense active participation. Absorptive capacity, in this sense, becomes a function of the same reputation-based system that governs scientific communities, rather than of pure firm-level cognitive mechanisms (Cockburn and Henderson, 1998).

3 Research context

Computer Science conference series

We focus our analysis on Computer Science (CS). Newell, Perlis, and Simon (1967) defined CS as “the study of computers and the major phenomena that surround them”. This research area has grown to include a variety of heterogeneous sub-fields spanning a wide spectrum between basic and applied research, from mathematic models or the physical properties of materials, to the development of hardware and software. At the same time, the economic relevance of this field of research is undeniable (Brynjolfsson and Hitt, 2003). Most of the main recent and ongoing technological revolutions stem from this field of research: from the transistor and the development of computers, to the advent of software, the Internet and Artificial Intelligence (AI).

CS is also an ideal setting for our study for pragmatic reasons related to the role of conferences within the field. Conference proceedings constitute a primary outlet for the publication of CS research results (Franceschet, 2010). They are peer-reviewed and the acceptance process is competitive.[5] Presenting the same manuscript at several conferences is rare, difficult due to the strict peer-review screening, which guarantees the originality of the work, and it is considered unethical.

As a consequence of their importance, conference proceedings in CS are also better represented in bibliographical databases as compared to other research fields, and curated information on conference series is available. This is convenient as it allows observing conferences on a large scale through bibliographic information. In addition, most conference organizers attempt to ensure the participation of authors, for instance, conditioning the actual publication of proceeding papers on the physical participation of at least one author to the conference.[6] This guarantees that the information obtained from conference proceedings will largely reflect the actual composition of active participants, or at the very least, of at least one member of a scientific team of coauthors.

CS is among the research fields with high private investments, and interactions and feedback loops between basic scientific advancements and insights from technological applications have been frequent (Nelson, 1962).[7] Scientific contributions are also often cited in patents (Ahmadpoor and Jones, 2017). However, the field is neither an outlier for the importance of scientific conferences for scientists (beyond the specific value of the publication of conference proceedings), nor for their relevance for the downstream industries (Cohen, Nelson, and Walsh, 2002), nor for the presence of firms at scientific events. Cohen, Nelson, and Walsh (2002) find from responses to the Carnegie Mellon Survey that conferences score similarly high across most industries as a channel of knowledge diffusion from public research to corporate R&D. Computers and semiconductors industry figure among them, but, if anything, they fall behind several other industries.[8]

We also explored descriptively the presence of firms at scientific conferences across different fields based on the affiliation information in conference proceedings in Scopus. Figure A-1 in the appendix presents the related results. In CS, 7.5% of all conference proceedings in Scopus are associated with firms. This share varies by fields: 9.7% in Physics, 11.3% in Engineering, 5.2% in Biochemistry/Genetics, 17.9% in Earth and Planetary Sciences.[9] With this observation in mind, we note that CS is not an outlier when it comes to the involvement of firms with scientific communities

[5] Based on publicly available information for a subsample of conferences, the acceptance rate is on average 21% at $A^{\star}$ conferences (N=333) and around 36% at B or lower conferences (N=988). Detailed data available upon request.

[6] See, for instance, the IEEE guidelines for conference organizers. The IEEE recommends the exclusion from or limitation of distribution of papers which were not presented at the conference. https://www.ieee.org/conferences/organizers/handling-nonpresented-papers.html

[7] High levels of industry participation in the field ML, in particular, has been sharply increasing in the last 10 years (Hartmann and Henkel, 2020).

[8] In computers and semiconductors, 37.9% and 48% of respondents indicated conferences to be important, respectively. Only publications and reports scored higher. This percentage is higher, for example, in petroleum (50%), drugs (64%), steel (54.6%), machine tools (45.5%), Aerospace (51%), and it is similar in several others (Cohen, Nelson, and Walsh, 2002).

[9] The methodology to identify conference proceedings associated with firms is an extension of the methodology discussed in section 4 to all conference proceedings in Scopus beyond CS. Note that the overall coverage of conference proceedings in Scopus is unknown. While we can assert that coverage for CS is fairly representative of all most relevant conference series (see section 4), we can not ensure this is the case for other fields. Descriptives are only indicative and should be taken with care.

at international conferences.

Firms' participation: authorship and sponsorship

Firms can actively participate in a conference having scientists presenting research results in conference proceedings and applying to be an official sponsor of the conference. To better understand the nature of these activities we gathered information on conference websites and we attended and interviewed participants at two major conferences: the European Conference on Computer Vision 2018 (ECCV, https://eccv2018.org/) in Munich, Germany and the Neural Information Processing Systems conference 2019 (NeurIPS, https://nips.cc/) in Vancouver, Canada. We interviewed more than 50 participants, between scientists and other representatives, of more than 20 firms and about 20 academic scientists. We discuss here the main findings. More details are given in Appendix D. A few footnotes in the paper also report key insights that relate to quantitative findings later presented.

The authorship of conference proceedings occurs normally as for other scientists, via the submission of proceedings that are selected in peer-review processes. The appearance of firm scientists as authors of proceedings largely coincides with their physical presence at the conference. Firm scientists present their work and normally interact with their academic and corporate peers. Almost all scientists of large firms we interviewed (with only one exception) declared to enjoy significant freedom to participate in conferences and that the acceptance of a proceeding constitutes a sufficient condition for all authors to have the support to participate. Scientists from medium to smaller firms also declared to enjoy similar conditions, only with more binding budget constraints. More generally, at the conferences we attended, the presence of firm scientists was staggering, and inspecting programs of these and numerous other conferences it appears evident as they also appear in scientific and organization committees, chairs and discussion roles.

Also for sponsorship, there is a limited number of slots available. Both firms and academic institutions can apply. The allocation happens on a first-come-first-served basis. Sponsoring fees range between a few thousand USD up to amounts in the order of $80'000, depending on the conference and sponsorship category.[10] The associated benefits go from: the more simple exposure of the company logo on the conference website, gadgets or venue; to the right to have exposition space at the conference; the possibility to submit applications for organizing talks, discussion panels, demos or workshops; and access to recruiting opportunities. Sponsors also have the right to one or more additional (non-presenters) entry tickets for individuals. In most cases, one or more employees among HR personnel and scientists are responsible to represent the firm at a booth. Here they provide information and disseminate material on the firm research and careers opportunities. Large sponsors also frequently organize workshops, tutorials, receptions and social gathering events.

To conclude, firms' participation, as we observe it, constitutes an actual firm-level investment

[10] For instance, the NeurIPS conference provides 5 sponsorship categories: Diamond ($80'000), Platinum ($40'000), Gold ($20'000), Silver ($10'000), Bronze ($5'000). In 2017, the conference attracted 84 sponsors with contribution fees totaling $1.76 million, an increase of 31.5% from the previous year, where 64 sponsors contributed for a total of $840'000). Source: https://medium.com/syncedreview/a-statistical-tour-of-nips-2017-438201fb6c8a

and a true engagement into the activities of scientific communities. The size of investments can vary substantially, based on the number of proceedings presented and sponsorship decisions (and the number of events per year). The cost of these investments goes beyond the simple attendance and sponsorship fees, including all costs of participation of scientists and personnel and preparation. [11] Passive attendance may occur but it did not appear to be the norm. Clearly, firms or firms' units opting for secrecy exist. These would be not present at conferences or more difficult to observe. However, the qualitative and quantitative information we collected show evidence of an intentional active participation of a large number of firms. Finally, specific firm-level processes have emerged from our interviews, complementary to the participation, such as the internal knowledge sharing of inputs from the conferences. These and other aspects are further discussed in Appendix D.

4 Data

We combine various data sources on conference series in CS and their participants between 1996 and 2015. Our primary objective is to cover a highly representative sample of all relevant conference series in CS with information relevant to our analyses. We make specific efforts and leverage data that allow us to reach sufficient disambiguation of conference series, firms and scientists. Table 1 lists the data we assemble and the corresponding information they provide. Appendix E offers a visual description of the connections between these datasets.

We obtain the central information on conferences from the Digital Bibliography & Library Project (DBLP). This is a database specialized on proceedings and publications in CS maintained at the University of Trier, Germany. We complement it with information from Web of Science (WoS) and Scopus, which contains the affiliation of authors, conference sponsors, citations and abstracts. The match between the two is highly precise, based on the DOI, when available, or key bibliographic information. We add information on conference series quality and CS research subfields from the Computing Research and Education (CORE) data, curated by the Computing Research and Education Association of Australasia. The CORE data are meant to classify all relevant conferences series in CS into the quality-rank levels $A^{\star}$, A, B and C and subfields based on experts' assessment. The match with CORE data is also highly precise, being done largely manually, with the support of text

[11] As an indicative example of the participation investment of a large firm, we take the participation of Google at the Neural Information Processing Systems (NIPS) conference in 2017 for which sponsorship costs are publicly available https://medium.com/syncedreview/a-statistical-tour-of-nips-2017-438201fb6c8a. Google figured as a second-tier sponsor of the conference, which corresponds to a price of $40'000. Seventy-five proceedings presented at the conference were authored by 86 distinct scientists. We can assume that 50% of them participated in the conference. We assume 5 participants from HR personnel, for a total of 48 participants. The conference took place in Long Beach, California and lasted for 6 days. We assume a travel cost of $130, a daily cost for accommodation and expenses of $200 per person, and an average yearly wage of $120'000, to divide by 260 working days. These assumptions are conservative, and neglect expenses related to the preparation and submission of conference proceedings and other general costs for the preparation of the material, the booth and conference activities. This sums up approximately to $260'000. Google, in 2017, participated, with varying intensity, to more than 160 conferences. Moreover, at the NeurIPS conference 2019 the large firms interviewed declared to have from 100 to up to 200 affiliated participants at the conference.

Table 1: Data sources.

Data source	Information
DBLP	Conference series, events, proceedings; author disambiguation
CORE	Conference series ranking; conference research sub-fields
WoS, Scopus	Affiliations; citations; conference sponsors; abstracts
SNPL data	Patent citations to conference proceedings
PATSTAT	Patents; applicant names
ICAO, BTS	Direct flight connections; Airport regions
ORBIS, GRID, EU Scoreboards	Firm names; ownership structure; industry information

Notes: Detailed version is available in the appendix, table A-14.

similarity indicators.

Data on airports and flight connections is obtained from the International Civil Aviation Organization (ICAO) and the US Bureau of Transportation Statistics (BTS). The ICAO data covers international flights, but domestic flights are not available. Since the US is one of the most important geographic areas for scientific activity in CS and flights are very important for US domestic travel the BTS data are an important complement. Both data sources come with a definition of market regions, usually the name of a city. We geolocalize all conference venues and scientists' affiliation and map them to airport regions in these data. We resolve cases of regions potentially assigned to multiple market regions choosing the busiest market region (highest passenger volume) within a radius of 100 km or the geographically closest airport beyond 100 km. We reach a total of around 1'100 relevant airport regions.

We match affiliations, sponsors as well as patent applicants with a custom firm database. Sources for data on firm names include ORBIS, the Global Research Identifier Database (GRID) and the EU scoreboards. From ORBIS, we take any firm associated with a patent as well as any firm in US and Germany (including subsidiaries), whether they patented or not. As such, it is a convenience sample, designed to identify all possible firms active at the conferences in our sample. The matching methodology is based on a supervised machine learning algorithm that combines, as input, information from search results for firm names in the search engine Bing (following the approach by Autor et al., 2020) and standard string similarity measures. We invest substantial additional manual post-processing efforts, especially to link entities that occur in multiple databases and whose identity and structure changes over time. We aggregate subsidiaries at the level of the corporate group.

We obtain patent level information from PATSTAT and match firms with patent applicants. PATSTAT contains patent information for all major patent jurisdictions worldwide, including information on inventors and applicants, their locations, and patent citations. However, citations to scientific articles are only available as strings within the broader field of non-patent literature (NPL) citations. One cornerstone in our data is an additional dataset where references to scientific articles in NPL citations are singled out and linked with bibliometric records in both WoS and Scopus (henceforth

SNPL data). The construction of this dataset is described in Knaus and Palzenberger (2018) and Poege et al. (2019).

We can claim that the data are largely representative of all relevant conference events in CS in our period of observation. In combination, WoS and Scopus cover the large majority of DBLP (90% since 1996). The CORE data do not cover a large number of small and less relevant conferences, with few corresponding proceedings each, which are dropped. However, 70% of WoS/Scopus-matched proceedings in DBLP are retained. Most importantly, the data cover 75% of all conference series listed in CORE and the highest share (80%) of CORE conferences not in our data are of the lowest quality rank, C. This implies that the data cover almost the entirety of top and medium ranked conferences and is biased against small and short-lived conference series of the lowest quality.

The sample is restricted to the period between 1996 and 2010 for practical reasons. Scopus is available to us from 1996 and we allow for a five-years window to observe dependent variables after the conference for a sufficiently long period, without truncation up to 2016 (which is the last year for which we dispose of citation information). This sample comprises 5470 firms, 7298 conference events pertaining to 1042 conference series, and a total of 612103 conference proceedings. In the remainder of the paper, we also present descriptive statistics limited to this sample. We cover in greater detail information on data sources and the construction of the dataset in Appendix E.

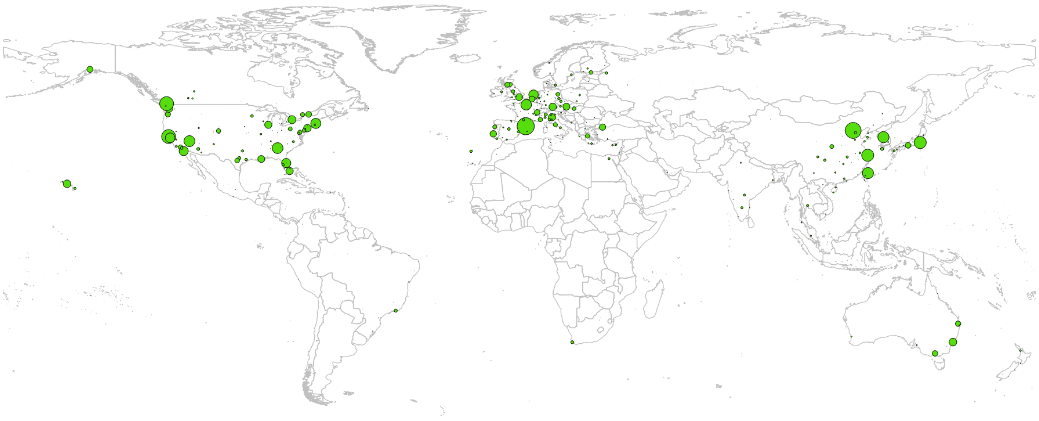
5 Descriptives

We present basic descriptive information on our data, with a focus on the characteristics of conferences in our sample and the level of participation of firms. Figure 1 shows the global distribution of conference locations. Conferences are distributed worldwide. The relevance of the traditional science regions of Europe and North America is as expected, but East Asia is a key region for conferences, too. It is interesting to note the relatively high frequency of locations that would not be normally encountered as major locations of scientists' affiliation, and are instead attractive conference venues. Table 2 reports the information on the sample of conferences. High-ranked conference series are fewer, but they are longer-lived and larger. Consequently, the total number of conference events in the sample is lower for A* and A conferences. However, the average number of events per conference series is higher for higher conference ranks: 12 (A*), 9 (A), 7 (B), 5 (C) (generally corresponding to an equal number of years). The average number of proceedings for conference event is 84, with up to 90 for highly ranked conferences, against 72 for C-ranked conferences.

Conference proceedings receive on average 3 scientific citations on a period of 5 years varying from 1.5 up to 9.8, for C and A* conferences, respectively. Patent citations are relatively rare, on average 0.15 per proceeding. It is perhaps more surprising to encounter also here a more than threefold increase, from 0.08 to 0.29, of citations from C to A* ranked conferences. When restricting to patent citations by firms in our sample, around 5.1% of conference proceedings are ever cited by a patent. For A^{*} conferences, this number rises to 8.2%.

Looking at the participation of firms, 88.3% of all conference events have at least one firm author

Figure 1: Location of conference events



Notes: Frequency-weighted conference airport regions are shown. The data is based on the estimation sample, counts for the years 1996-2010 for conferences where at least one firm was present are aggregated.

Table 2: Conferences information by rank

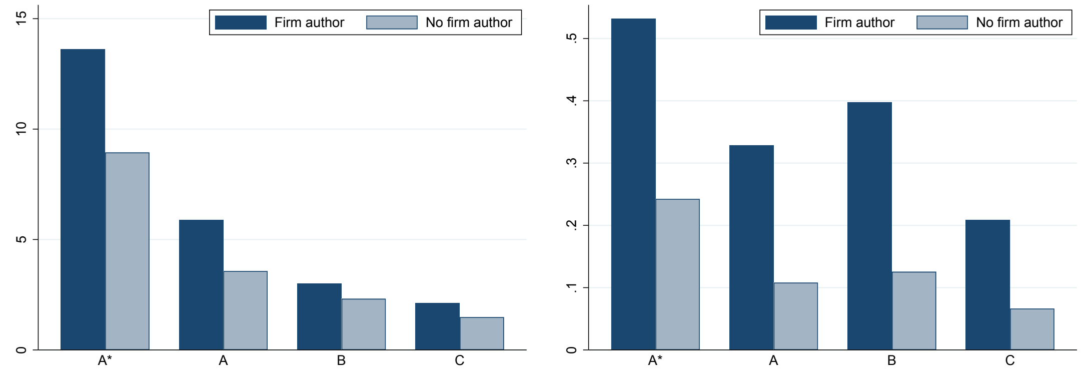
Rank	$A^{\star}$	A	B	C	Total
Conference series - Total	68	199	355	415	1037
Conference events - Total	824	1875	2597	2002	7298
Firms participating - Avg. n. per conference event	7.81	6.12	6.30	4.15	5.84
Firms sponsoring - Avg. n. per conference event	7.19	5.49	5.68	3.55	5.22
Proceedings - Avg. n. per conference event	89.17	84.93	90.42	72.21	83.87
Share of firm-authored proceedings - (%)	17.68	10.78	10.93	7.43	10.87
Science citations (5y) - Avg. n. per proceeding	9.77	3.82	2.39	1.53	3.45
Patent citations (5y) - Avg. n. per proceeding	0.29	0.13	0.16	0.08	0.15

Notes: Data for years 1996-2010 is used.

and 10.9% of all proceedings have at least one firm as author affiliation. 26.6% of conferences, corresponding to 29.6% of proceedings, have at least one firm as sponsor. From Table 2, the average number of firms participating as authors' affiliations at conference events and as sponsors is 5.8 and 5.2, respectively. It is interesting to note that the intensity of this participation increases strongly with the quality-rank of conference series. In percentages, in $A^{\star}$ conferences, 17.7% of contributions are by firms. At levels A and B, around 10.8% of contributions are by firms and at C-level conferences, only 7.4% of contributions are by firms. Of $A^{\star}$ conferences, 39.8% are sponsored by at least one firm, down to 22.2% at C-level conferences.

While comprising only 10.9% of the total, firm authored proceedings stand out in terms of quality. Figure 2 (a) shows the average count of scientific citations to proceedings with and without firm affiliated scientists, by conference rank. Firm proceedings are of exceedingly high quality, for any level of conference rank. For instance, within $A^{\star}$ conferences, firm proceedings receive on average 5 more forward citations within five years compared to non-firm proceedings, that receive 9 citations

Figure 2: Citation counts by type of authors' affiliation and conference rank



(a) Science citations

(b) Patent citations

Notes: Counts forward citations by any CS paper or proceeding (2a) or by any patent family (2b), by conference quality and by author status. A citation window of five years for conference proceedings published in 1996-2010 is used. Cf the overall averages in table 2.

on average. The results are purely descriptive but suggest that firms have a scientific impact and gain considerable attention within the scientific communities where they participate. [12] At the same time, figure 2 (b) shows that firm authored proceedings are also cited in patents at a much higher rate. The number of patent citations decreases for lower conference ranks, but less sharply than scientific citations.[13]

Table 3: Firm information

Variable	Mean	SD	25%	50%	75%	90%	99%
Firm Scientists	18.40	155.70	1	2	6	17	315
Conference participations	8.15	61.20	1	1	3	8	139
Conference sponsorships	0.86	9.87	0	0	0	1	11
Conference proceedings	13.72	163.82	1	1	3	9	205
in collaboration w. academics	7.14	82.13	0	1	2	5	105
in collaboration w. other firm	1.53	14.27	0	0	1	2	25
Firm Patents citing proceedings	8.24	100.02	0	0	0	3	160
Observations	5224						

Notes: Descriptives at the firm level for the 1996-2010 period. Scientist number is the maximum yearly number in this time period. Conference participations/sponsorships count firm-conference events. Conference proceedings instead count the total number of proceedings of that firm, first overall and then broken down by type. Lastly, the number of patent families citing a proceeding is counted.

[12] This result resonates well with qualitative evidence from our interviews. Most of the firms declared to have internal peer-review systems to guarantee that the work they present is of above-average quality within the events where they participate in. Some firm scientists have also expressed the idea that, differently to academia, they had no pressure to publish and they would focus on presenting work that they deemed of high impact (from a scientist of a startup in the field of ML: "We are not in a publish or perish mode").

[13] In regression analyses in the appendix table A-3, we test the robustness of these correlations to conference series and also conference events FE. The analysis confirms that also within conference series and single conference events, proceedings authored by firms receive more citations. This is particularly the case when the firm is also a sponsor of the event.

Finally, Table 3 presents information at the firm level. On average, about 18 different scientists have worked for a firm in our sample and authored at least one proceeding. They have authored about 13 proceedings presented at 8 distinct conference events. Firms have then sponsored 0.86 conference events and filed 8 patents citing one or more conference proceedings, on average. Interestingly, a relatively large share of proceedings (7 out of 13.7, on average) is coauthored with academic scientists. On the other hand, collaborations with other firms are rarer (1.5 proceedings). Still, a significant share of proceedings (about 45%) is authored exclusively by scientists affiliated to one single firm. Importantly, the distribution of these firm-level indicators is extremely skewed. Firms below the median had only 2 active scientists or less, they have authored only one proceeding, they never sponsored an event, and have no patents with citations to proceedings. The large share of activity is observed above the median and, even more, at the top quarter, top 10% and top 1% of the sample (roughly 1250, 500 and 50 firms, respectively).

We present in Appendix B the variation over time of some key descriptive statistics. Overall, not surprisingly, the number of conferences and conference proceedings grew substantially (A-2a). The fields of Artificial Intelligence and, secondly, Information Systems have grown the most. The increase in the number of conferences in non-US locations has been more prominent (A-2b). In 1996, half of all conferences were taking place in the United States, by 2015 it was about a quarter. In terms of composition, the figures discussed above on the level of participation of firms have remained fairly constant over our period of observation. Interestingly, the share of proceedings in collaboration between academia and firms has increased steadily, but not collaborations between firms that instead remained rare and constant in share.

6 Econometric strategy

6.1 Econometric model

Our empirical analysis focuses on the effect of the participation in the same scientific events on knowledge diffusion from scientists to firms. A first challenge is that we only observe realized conference participations. Second, both participation of firms and scientists in conferences is likely endogenous. Firms choose consciously in which conferences to participate, favoring, as we show descriptively, high-quality conferences. Similarly, both academic and firm scientists likely select conference events based on their interests, the quality of the conference, and expectations regarding the participation of other scientists and institutions. A simple comparison of knowledge flows from scientists to participating firms versus knowledge flows to absent firms would not suffice. It would be impossible to distinguish the effect of participating in the same conferences from the unobserved factors that determine the selection of a conference in the first place. To test our hypothesis we must address this fundamental inference problem.

We work with a dataset of pair-level observations of firms and conference proceedings. We maintain all firm-proceeding pairs for proceedings presented at conferences where a firm participated.

We then add, for each conference, a counterfactual group only of proceedings presented at similar conferences. We create strata of conferences in the same year, rank, subfield, and within the same size and forward-citations count categories. To create these categories, we coarsen the conference size into three categories (≤25%, ≤50%, >50% of the largest conference within the group) and forward citation counts using a median split for A* and A-ranked conferences and quartiles for B and C-ranked conferences. We retain up to two other conferences selected randomly within the same strata. The results are robust to selecting only one or more than two conferences. We verified that the average difference for any observable between matched conferences cannot be distinguished from zero.[14]

The matching process generates variation in the data between pairs of firms and proceedings presented and not presented at the conferences where the firms participated. In this way, the counterfactual sample is a sample of proceedings also aligned to the revealed interest of firms and scientists. The proceedings not presented at the same conferences also serve as a counterfactual to the observed participation of scientists authors of the proceedings. Our approach only requires the assumption that, a priori, the scientists could have presented the proceedings at a matched counterfactual conference selected, with some probability. We do not require the probabilities of participating to the actual or counterfactual conference to be the same or the choice to be random. In other words, the matching procedure does not solve the problem of the endogeneity of participation. Other unobserved factors, besides those considered, that correlate with both the decision to participate as well as the likelihood of knowledge diffusion, are likely to exist.

To address endogeneity we employ econometric models that isolate exogenous variation in the probability of participation of scientists. The availability of direct flights to the conference venues from the location of scientists provides such variation. Direct flights tend to reduce costs, travel time and eliminate layovers. The presence of direct flights is also more likely associated to airport pairs where competition between airlines is higher, therefore reducing the price also of other options. This affects the general costs of transportation and thus the probability of participation. Flight connections have been used before as proxy for the cost of physical individual interactions within firms (Giroud, 2013), between scientists (Catalini, Fons-Rosen, and Gaulé, 2020) and as determinant of cities economic growth (Campante and Yanagizawa-Drott, 2017). Our study differs as we observe the endogenous variable - participation in conferences - and use direct flights as an instrumental variable. We represent our empirical setup graphically in a stylized scenario in figure 3.

We implement a two-stage regression model. The main endogenous variable of interest is labeled *Participation* and is a dummy variable that takes value 1 if the proceeding p was presented in a conference series c where firm f participated. This is evidence that firm f and at least one author of conference proceeding p has participated in the same conference. The variable takes value 0 for proceedings matched to conference c that were actually presented at a conference where firm f did

[14] Results of this analysis are presented in table A-1. Note that each conference in the sample is randomly pared to one or more other conferences within the same sample (and according to the matching strata). Consequently, this analysis is not a test of equivalence of matched conferences. It is purely aimed at excluding a malfunction of the matching algorithm and random selection.

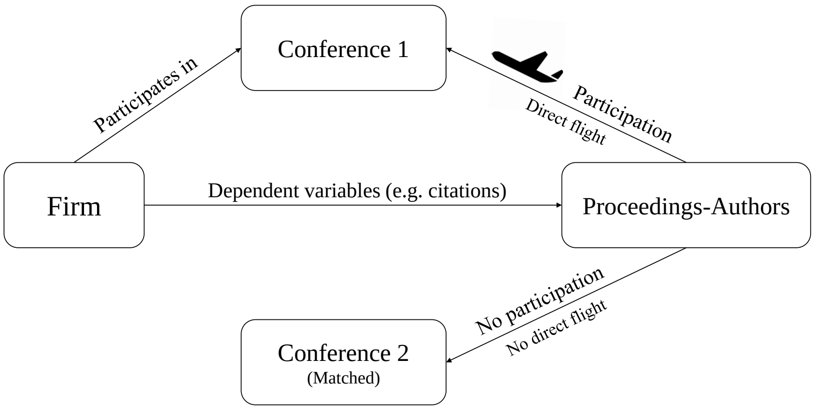


Figure 3: Stylized empirical setup

not participate. The first stage equation 6.1 models the probability of *Participation*, so defined, as a function of the existence of a direct flight between the location of the authors of p and the venue of conference c, in the year that p is presented. The variable *Direct Flight* is equal 1 if a direct flight exists. X_{fpc} is a vector FE controls and control variables. [15]

First stage:

$$P(\text{Participation})_{pcf} = \beta_1 \text{Direct flight}_{pcf} + \beta_2 X_{pcf} + u_{pcf} \tag{6.1}$$

The second-stage equation (6.2) models the probability of knowledge diffusion from the conference proceeding p and its authors to the firm f as a function of *Participation*. We use as dependent variables a set of indicators of knowledge diffusion, better defined in the next section 6.2.

Second stage:

$$P(\text{Knowledge diffusion})_{pcf} = \gamma_1 \text{Participation}_{pcf} + \gamma_2 X_{pcf} + \epsilon_{pcf} \tag{6.2}$$

The identification assumption in this model relies on the exogeneity of the availability of direct flights between the specific location of scientists and the venues of conference series, with respect to the probability of knowledge flows between scientists and firms. A specific advantage of our setting is that the pairs of scientists and conference locations, in the large majority of cases, do not coincide with the pairs of scientists and firms locations. Possible preexisting relationships between firms and scientists in a given location would hardly influence the connectivity with conference venues, if these do not coincide. Moreover, this feature strengthens the credibility of the exclusion restriction - i.e. the assumption that the instrument affects knowledge flows exclusively via the participation to the same conferences - because new direct flights would not imply, in general, increased direct connectivity between firms and scientists.

Airlines' and conference organizers' decisions depend on several factors. New airline routes

[15] The conference proceedings p are nested in years so that we do not have another index for years. We omit the intercept in the specification because it is always collinear to the FE controls included in X_{fpc}.

are likely driven by broad market trends and regulations (Campante and Yanagizawa-Drott, 2017). The location decisions of conferences series are driven by the general attractiveness of venues but are likely independent from the specific pair-level probability of interaction between scientists and firms in a specific year. Conference locations are scheduled often years in advance and organizers choices are primarily constrained by budget considerations and the need for adequate venues in terms of size and surrounding facilities. Some conferences are static, while others show rather erratic patterns of mobility, besides the preference for attractive locations. As long as the decision factors of airlines and conference organizers are unrelated to the pair-level probability of knowledge flows between scientists and firms, they are not a concern. However, (omitted) factors affecting both the existence of direct flights and the probability of knowledge diffusion may still bias the results. In our specification, we can account for these factors at the level of the conference series, the firm and the location of scientists. In this, our approach is akin to previous studies, in particular Giroud (2013).

Levels FE. Without additional controls, we would have to worry, for instance, that the most innovative firms, the most productive researchers, and most highly ranked conferences are likely located in regions with better airline connectivity. The estimates, rather than an effect of participation, would reflect the fact that these firms and scientists will more likely cite (or receive citations) and participate in the best conferences. Accordingly, we include FE controls for all the main levels of observation: conference series, region of origin of researchers interacted with CS sub-fields, years. Conference series, region of origin of researchers FE are necessary to control for any constant characteristics related to these levels of observation that can correlate with access to direct flights. We include the interaction between regions and sub-fields as FE to account for regional specialization. Year FE account for general time trends in the data. Firms fixed-effects are also included, nested in firm and scientists location pair-FE, discussed next.

Firm and scientists location pair-FE. Another concern is the possible correlation between direct flights to conference venues and the connectivity between pairs of scientists and firms locations. This would be the case for pairs of firms and researchers in the same or proximate regions. A new airline, for instance, would likely increase the number of direct flights to conference locations, but also between all locations within the same geographic area where it operates. Close firms and researchers will also more likely participate in the same conferences and, at the same time, exchange knowledge directly because of their proximity. We control for firm and scientists origin pair-FE, to control generally for all pair specific features between the firm and the location of scientists. This includes geographic distance, location in the same countries or regions, spoken language similarity between different countries, etc. Moreover, these pair-FE control also for possible specific connections of the firm with the scientists' locations that we cannot observe (e.g. presence of subsidiaries).

Firm-level shocks: firm and year pair-FE. To control for firm-level shocks, we are able to control for firm and year pair-FE. An increase in innovative productivity specific to one or more firms and in a specific period would affect their propensity to participate to conferences, particularly of high quality, and, at the same time, their capacity to absorb external knowledge. If such trends are indeed firm and time-specific, level-FE would not suffice. Firms with increased productivity would suddenly participate in the best conferences that may be better connected via direct flights to scientists' locations. Positive estimates may partly reflect the fact these firms may absorb more likely scientific knowledge, independently from conference participation. Firm and year pair-FE eliminate this type of concerns.

Scientists' location-level shocks - scientists location and year pair-FE. Finally, we account for time-specific shocks at scientists' locations level. Economic and innovation trends may be region and time-specific. Better infrastructures, including transportation networks, would normally follow or precede such trends. In this case, despite the use of time-invariant FE, the presence of direct flights may correlate with the quantity and quality of scientific activities in a region, in specific years. Scientists that increase their participation to conferences, thanks to the higher attractiveness of their regions for airlines and conference organizers, may be also those more likely cited in a given period, regardless of their actual participation. We control for scientists location and year pair-FE to fully absorb this variation.

We model both stages as Linear Probability Models (LPM). The use of LPM eases the interpretation of the coefficients, that can be interpreted as changes in percentage points in probabilities. Non-linear probability models and count models are hardly applicable due to our sample size and the use of high-dimensional FE. Due to the structure of our data and analyses, we have to account for serial correlation and within groups correlation of the standard errors. In our main specification, we cluster standard errors at the level of region of origin of scientists. In section 9 and in the appendix we test additional specifications to address other concerns: additional FE controls; continuous outcome variables; alternative clustering of standard errors; OLS specifications; falsification test analyses.

6.2 Definition of variables

Knowledge diffusion

Our main dependent variables capture knowledge diffusion from scientists to firms as measured by citations from firms to proceedings (Jaffe, Trajtenberg, and Henderson, 1993; Roach and Cohen, 2013). We look both at scientific citations from firm publications and from non-patent literature (NPL) citations in firms' patents. Naturally, scientific citations represent knowledge diffusion to firm science activities, while patent citations to innovation activities. Patents are imperfect proxies for innovation, but they represent, differently from scientific publications, actual intellectual property assets. All the more, patents with citations to science are consistently found to be highly valuable (Ahmadpoor and Jones, 2017; Poege et al., 2019; Schnitzer and Watzinger, 2019). Science and innovation are also often distinct operations of independent organizational units, especially in medium to large firms. Hence, the importance of looking at both dimensions.

Some have expressed concerns on the validity of citations as indicators of knowledge flows. Patent citations serve legal functions and are often added by attorneys and examiners (Alcacer and Gittelman, 2006; Thompson and Fox-Kean, 2005). Science citations are more normally associated with the actual recognition of knowledge sources by scientists, but they may also reflect strategic behavior or the saliency of a reference (Teplitskiy et al., 2020). Precisely because the data generation process for the two indicators is so radically different, finding an effect on both would be reassuring. Moreover, Roach and Cohen (2013) have argued that NPL citations to science reflect more likely actual knowledge flows compared to patent to patent citations. Some concerns are also ruled out under our empirical setting. For instance, in the absence of actual knowledge flows that make an invention similar to a proceeding, it is unlikely that examiner citations would be influenced by the participation of firms to a conference per-se. In section 9, we address further this issue using alternative dependent variables based on text similarity.

Science cit (present) and *Patent cit (present)*. We denote with the variable name *Science cit.* scientific citations and with *Patent cit.* patent citations. We look first at citations to the proceedings at the conference. Specifically, the label *present* refers to citations to the same conference proceeding p in the firm-proceeding pair observation. All citation-based variables are defined as dummy variables equal to 1 if at least one citation is observed, and 0 otherwise. To avoid truncation in the latest years, we restrict the sample to conferences up to 2010 and we look at citations within a 5 years time window after each conference.[16] For patents, we count the years of delay based on their priority year.

Science cit (past) and *Patent cit (past)*. We look separately at citations to previous publications of the scientists, labeled as *past*. These are citations to other publications from the same authors of the proceeding p, published in the same or the previous five years. The window of time considered after the conference remains also 5 years, as discussed above.

Control variables

The main endogenous variable of interest *Participation* and the instrumental variable *Direct flight* are discussed previously (see 6.1). We use additional independent variables as controls. We do not consider these controls as necessary for identification. It is meaningful to conceive the determinants of airline routes to conference venues as being a function of broader regional determinants. Conditional on the FE-controls in our models, we deem unlikely that there is any residual correlation between the presence of direct flights and proceeding level or firm-proceeding pair-level variables. We include these variables for robustness.

Conference distance controls. We control for indicators of geographic distance between scientists´ location and conference venues that, in the first stage, may influence the probability of participation. In particular, we control for the logarithm of the geographic distance between the scientists' locations and the conference venues (*Conference distance*). Geographic distance is measured as the minimum great-circle geodetic distance between the conference venue and the location of the authors of the

[16] For patent citations this may be insufficient to eliminate truncation. Citations are added to patent families over time in subsequent publications of the patent (e.g. grant publication and international filings), both by examiner and applicants. As grant lags of several years are not uncommon, many citations may remain unobserved for the latest conferences in our sample. We ensured that this is not an issue by running regressions for a sample of conferences up to 2008, finding equivalent results.

proceeding p. We add discrete measures of distance. First, a dummy equal one if the conference takes place in the same region of the scientists, defined by regions served by the same airports (*Same region*). Second, a dummy equal one for locations within the same countries or the same state in the US (*Same state*)[17]. Notably, these controls also exclude from the identifying variance the possibly also exogenous variance deriving from domestic and short-distance direct flights. This is an additional guarantee against the concern that conference venues and specific scientists may be co-located, biasing the results.

Firm-proceeding controls. At the firm-proceeding pair-level, we control for indicators of similarity and relevance to the firm R&D of the proceedings and the previous research of scientists. These controls are motivated by the second-stage, as likely predictors of knowledge diffusion. *Science citations (L)* captures the presence of citations from proceedings of the firm in the previous 5 years up to the year of the conference to previous publications of the authors of the proceeding p. Similarly, *Patent citations (L)* capture citations from patents of the firms. The two variables can be seen as the analogous of lagged dependent variables. We use, instead, a measure of the research similarity of the firm to the focal conference proceeding p, looking at the text-similarity (of title and abstracts) with firms' proceedings in the year before the conference, in the same CS sub-field.

Mechanisms

Collaboration. Our exploration of mechanisms departs from considering actual social interactions with scientists as a key channel of knowledge diffusion. In particular, we want to estimate whether participation affects the probability of collaboration with scientists. We cannot capture all forms of collaboration, especially if informal or unsuccessful. However, we can observe scientific collaborations that lead to future coauthored publications. We capture the presence of such collaborations with the variable *Collaboration* which is a dummy equal to 1 if in the 5 years following the conference at least one scientist author of proceeding p is found to co-author at least one publication with the firm. We use this as an additional dependent variable.

Hiring. The second channel of particular interest is mobility. We capture this aspect with a variable *Hiring*, which is equal one if at least one scientist author of the proceeding p (not affiliated to the firm f at the time p is presented) is found to publish with the firm f as affiliation in the 5 years following the conference. This is an imperfect proxy because hiring is observed exclusively if the scientist publishes. This is likely if the scientist is hired in a research unit but less so in a product development unit. However, this variable is not meant to perfectly measure hiring. It rather serves the purpose to assess whether our results can be explained by hiring, rather than by knowledge flows and collaborations with external scientists. We use also this variable as a dependent variable.

Participation intensity - N.Proceedings. We study the role of firm participation intensity. The variable *Participation* captures the participation of firm and scientists to the same conference, independently from the modes of participation of the firm. We interact this variable with the two

[17] The latter control variable is also motivated by the fact that while we have complete flight information for US domestic flights, we mostly observe only international flights for other countries

variables indicating more intense participation. First, *N.firm-proceedings* is the number of proceedings presented at the conference by the firm‘s scientists. The variable captures the level of presence of the firm in the scientific program of the conference. Because the variable distribution is highly skewed and particularly sparse we aggregate the values of 3 and 4 proceedings together and we censure the variable at the value of 5 for firms with 5 or more proceedings at a single conference event. The median number of papers at a conference is 1, at a mean of 1.6. Only the top 4.9% of our estimation dataset has 5 or more proceedings presented by one same firm.

Participation intensity - Sponsorship. Second, we use *Sponsor*, a dummy equal 1 if the firm is a sponsor of the conference. Sponsorship is a financial contribution and a specific investment for increasing reputation, that can be in principle independent from scientific contributions. We distinguish the case of firms that exclusively sponsor a conference event, without also having papers presented (*Sponsor-only*) and the case of firms doing both (*Sponsor-Proceedings*). In our estimation dataset, 6.3% of firm attendances come with sponsorship. Of those, in 30.0% of the cases, firms are both participating and sponsoring.

Firm size. Finally, to proxy size of research investments, we rank firms within each year by the number of active scientists they employ, as proxy for the size of research units. For this, we build an affiliation panel for each scientist. When there are multiple affiliations within a year, we use fractional counts. When there is no publication in a year, but in earlier and later years, we use linear imputations. We use all types of publications to build the panel, including journal articles, books and editorials. We use firm size ranks instead of fixed size groups because, over the years, the number of items covered in the bibliometric databases has increased substantially (see figure A-6). We look at the top 5, top 50 and other firms' groups, within each year. Overall, 9.6% of firms in our full estimation dataset appears at least for one year in the top five firms and 30.3% in the top six to fifty.

7 Main results

7.1 First stage: participation

The first stage regression results are reported in Table 4. For the sake of brevity, we present here only the first-stage result corresponding to our first dependent variable, *Scientific cit. (present)*. The sample for different dependent variables varies minimally due to some observations being invariant within FE-controls, but otherwise the first-stage estimates remain almost identical. In this and all tables for the main dependent variables, we present a series of 3 specifications adding controls gradually. The first column for each dependent variable includes exclusively level FE controls, for conference series, for scientists locations and subfield, firms and scientists' locations pairs, and years. The second column adds time specific FE: firm and year pair-FE, scientists location and year pair-FE. The last column is the full specification including all FE controls and the additional control variables described in section 6.2.

Table 4: First stage - the effect of *Direct flight* on *Participation*

	Science cit (present)		
	(1)	(2)	(3)
Dep. Var.	Participation	Participation	Participation
Direct Flight	0.056^{***}	0.059^{***}	0.030^{***}
	(0.006)	(0.006)	(0.005)
Firm-proceeding controls			
Science citations (L)			0.118^{***}
			(0.003)
Patent citations (L)			0.055^{***}
			(0.004)
Research similarity (L)			0.958^{***}
			(0.025)
Conf. distance controls			
Conference distance			-0.039^{***}
			(0.003)
Same region			-0.160^{***}
			(0.035)
Same state			0.131^{***}
			(0.015)
Conf Ser FE	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes
Year FE	Yes		
Origin × Firm FE	Yes	Yes	Yes
Year × Origin FE		Yes	Yes
Year × Firm FE		Yes	Yes
R^2	0.297	0.323	0.346
Observations	5126376	5126273	5126273
Number clusters	1124	1114	1114
DV cond. mean	0.512	0.512	0.512

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). First stage results for other second-stage variables are very similar, results are available upon request.

The results show a strongly significant effect of the instrumental variable *Direct flight* on *Participation*. In all specifications, we find a highly significant and positive coefficient. The magnitude is economically meaningful, implying that the existence of a direct flight leads to an increase in the probability of a proceeding being presented at a conference of about 5.9 p.p. or 3 p.p in the full specification. Since this probability is 51.2% in the sample, this corresponds to about a 12 or 6% increase in probability, respectively.[18] We report the Kleibergen-Paap rk Wald F statistic. The F-test

[18] Recall that we match each actual participation with up to two counterfactual conferences. The average of the variable *Participation* remains higher than one third because in many cases the matching process narrows down the search to the one only alternative conference and because in some cases the matched conference is also a conference where the firm participated.

value on the excluded instrument always exceeds a value of 20 and is often substantially higher, depending on the specification. Note that some controls (*Same state* in particular) absorb part of the same variation of the instrumental variable, so, the lower coefficient and significance of *Direct flight* is expected in the third specification.

The control variables estimates show rather predictable correlations. The probability of participation decreases with *Conference distance* and increases for *Same state*. The negative coefficient on *Same region* is more surprising. However, this is conditional on the other two distance controls. The coefficient subtracts to the positive effect of *Same state* and geographic proximity. The result probably reflects a preference to travel outside one own region, conditional on transportation costs being low. The results on the other control variables seem to suggest that proceedings from authors previously cited (*Science citations (L)* and *Patent citations (L)*), that are similar to the firm research (*Research similarity (L)*) are in general more likely to participate.

Appendix A.2 presents an heterogeneity analysis of the effect of *Direct flight* which is informative to understand the population relevant for the Local Average Treatment Effect (LATE) estimated. We find predictable variation. The effect of *Direct flight* is significant for all conference ranks in the specification without distance controls, and is stronger for medium and low ranked conference series. The introduction of distance controls reduces the power of the instrument particularly for highly ranked conferences. However, the effect of *Direct flight* remains significant, and strongly significant and larger in magnitude for conferences at long geographic distances, regardless of their ranking.

Finally, in Appendix A.1 we propose an event study analysis, at the scientists' location - conference series pair level of analysis. This analysis demonstrates the lack of any pre-trend in the probability of participation of scientists prior to the first time that a direct flight connection to participate to a conference series is available. The probability of participation increases immediately at the event time and slightly increases in the following years if the flight connection is maintained.

7.2 Second stage: knowledge diffusion

We first present, in table 5, the result of the second stage regressions for *Science cit (present)* and *Patent cit (present)* (the probability of scientific citations and patent citations to the focal proceedings). The coefficients indicate the change in probability in percentages points (p.p.) for proceedings presented in the same conference where the firm participates (*Participation* equal 1). Columns 1 to 3 present results for *Science cit (present)* as dependent variables, columns 4 to 6 for *Patent cit (present)*. For both dependent variables, we deploy a series of model specifications, as for the first-stage results, gradually including FE and control variables. Specifically, columns 1 and 4 include only main FE (conference series, scientists locations and subfield, firms and scientists' locations pairs, and years), column 2 and 5 add year specific FE (firm and year pair-FE, scientists location and year pair-FE), and columns 3 and 6 add control variables (described in section 6.2).

The second and third are the favorite specifications. However, the coefficient magnitude varies minimally between the first and second specification, also in all other results. We find significant results, at the 1% level, for the effect of *Participation* on *Scientific cit. (present)*. The magnitude of the coefficients varies more, from 0.013 in column 2 to 0.021 in the specification in column 3, corresponding to 1.3 and 2.1 p.p. change in the probability of observing a citation. This change in magnitude is found also in most other analyses and can be attributed to a change in the LATE estimation. As discussed in section 7.1, and demonstrated in Appendix A.2, the instrument has a much larger effect on low-ranked conferences when we introduce conference distance control variables. Therefore, a possible explanation is that the effect of participation in lower-ranked conferences is larger, because proceedings at this conferences, and their authors, would likely attract less attention or be less likely met, unless encountered at a conference.

We use the sample mean average conditional on actual participation as benchmark (the average probability of citation for proceedings at the conference where the firm participates). Compared to this baseline average probability of citation of 1.0%, the effects are quite substantial. On the contrary, the coefficient for the effect on *Patent cit. (present)* is small, at 0.001 and not statistically different from zero across all specifications. In other words, we do not find an effect of *Participation* on the probability that conference proceedings presented at a conference are cited in patents.

The correlation with the control variables is meaningful, showing that proceedings that are similar to the firm recent research or from authors that have previously been cited by the firm are more likely to be cited again, both in firm publications and patents. We can use these estimates as a benchmark for the effect of *Participation*, noting that it has an almost comparable magnitude to *Scientific citations (L)*; this is, proceedings of authors already cited in the past by the firm are more likely to be cited again, but the effect *Participation* is of the same order of magnitude as this correlation.

In Table 6 we move on to the results for the variables *Science cit (past)* and *Patent cit (past)*. These show the effect of *Participation* on the probability of citations to other publications published by the same authors of the focal proceedings up to 5 years before the conference. We again include FE and control variables gradually as just discussed for Table 5. Point estimates for the effect on *Science cit (past)* are 6.9 p.p., in column 1 with basic FE control, to 0.057 in column 2, with significance levels at 5%, and 11.3 p.p in column 3 with significance level at 1%. This is again a large increase compared to the conditional sample mean, 15.8%. The effect on patents citations, from the results in column 4 to 6, is now highly significant and large in magnitude, corresponding to 4.7 in the more basic specification (column 4), 4.8 in column 5, and 9.8 p.p in column 6.

The results relative to the control variables are again predictable, implying, for instance, that proceedings previously cited by a firm are more likely to be cited again. To use these figures as a benchmark for the coefficients of interest, we can say that *Participation* increases the probability that the scientists are cited after a conference by about one-fourth of the probability increase associated with the scientists having been cited already before the conference. This ratio is a bit less than a half comparing the effect of *Participation* on *Patent cit (present)*, relative to the correlation with *Patent cit (present)*. This highlights that these are economically meaningful effects.

Overall, the results in table 5 and 6 demonstrate a strong effect of *Participation* on citations. The comparison between the different dependent variables allows some considerations. The effect

Table 5: The effect of *Participation* on citations to proceedings at the conference

	(1) Science cit (present)	(2) Science cit (present)	(3) Science cit (present)	(4) Patent cit (present)	(5) Patent cit (present)	(6) Patent cit (present)
Participation	0.013^{***}	0.013^{***}	0.021^{***}	0.001	0.001	0.001
	(0.004)	(0.004)	(0.007)	(0.001)	(0.001)	(0.003)
Science citations (L)			0.029^{***}			0.003^{***}
			(0.001)			(0.000)
Patent citations (L)			0.008^{***}			0.003^{***}
			(0.002)			(0.001)
Research similarity (L)			0.025^{***}			0.007^{**}
			(0.006)			(0.003)
Conf. distance controls	No	No	Yes	No	No	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes			Yes		
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE		Yes	Yes		Yes	Yes
Year × Firm FE		Yes	Yes		Yes	Yes
R^2	0.076	0.082	0.083	0.024	0.030	0.031
Observations	5126376	5126273	5126273	5126376	5126273	5126273
Number clusters	1124	1114	1114	1124	1114	1114
DV cond. mean	0.010	0.010	0.010	0.002	0.002	0.002
F (First)	81.1	88.7	31.6	81.1	88.7	31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Citations of firm science (columns 1-3) and firm patents (columns 4-6) in the subsequent five years towards the focal proceeding are analyzed. The dependent variables are 1 if at least one citation occurred. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.

on scientific citations is significant for both proceedings at the conference and previous proceedings of scientists and, if anything, it is slightly stronger for the former. On the contrary, the effect on patent citations is exclusively significant for the latter. This can derive from the difference between science and innovation activities within firms. In particular, we may expect scientists participating in conferences to be the same that work on firm science and can immediately build on new knowledge inputs in upcoming publications. This may not be the case for innovation activities that are probably performed by distinct organizational units and require farther development to reach outputs such as a patented invention. The lack of significance for patent citations to focal proceedings may also be due to patent citations being a noisier indicator.

For the purpose of our investigation, we note that if we had found an effect exclusively for proceedings at the conference we would have concluded that the timely screening of information at the conference was the main mechanism explaining citations. It remains possible that proceedings

Table 6: The effect of *Participation* on citations to previous proceedings of scientists

	(1) Science cit (past)	(2) Science cit (past)	(3) Science cit (past)	(4) Patent cit (past)	(5) Patent cit (past)	(6) Patent cit (past)
Participation	0.069^{**}	0.057^{**}	0.113^{***}	0.047^{***}	0.048^{***}	0.098^{***}
	(0.028)	(0.025)	(0.041)	(0.015)	(0.013)	(0.024)
Science citations (L)			0.394^{***}			0.158^{***}
			(0.006)			(0.003)
Patent citations (L)			0.184^{***}			0.195^{***}
			(0.006)			(0.006)
Research similarity (L)			0.240^{***}			0.012
			(0.043)			(0.024)
Conf. distance controls	No	No	Yes	No	No	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes			Yes		
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE		Yes	Yes		Yes	Yes
Year × Firm FE		Yes	Yes		Yes	Yes
R^2	0.304	0.318	0.372	0.176	0.187	0.192
Observations	5126376	5126273	5126273	5126376	5126273	5126273
Number clusters	1124	1114	1114	1124	1114	1114
DV cond. mean	0.158	0.158	0.158	0.050	0.050	0.050
F (First)	81.1	88.7	31.6	81.1	88.7	31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Citations of firm science (columns 1-3) and firm patents (columns 4-6) in the subsequent five years are analyzed. Citations towards publications by the proceeding authors in the five years before the conference are considered. The dependent variables are 1 if at least one citation occurred. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.

at the conference serve as pointers to previous proceedings of the same authors. However, we argue that the effect on “citations to the *past*” is at least supportive of the hypothesis that actual social interactions with scientists are the central channels of diffusion. In other words, if the scientists rather than the proceedings are the main source of knowledge, the fact that other recent proceedings and those presented at a conference are similarly cited is less surprising.

8 Exploration of mechanisms

8.1 Collaboration and Hiring

The following analyses seek to provide direct evidence on the effect of *Participation* on the *Collaboration* with and *Hiring* of scientists. While we do not capture these dimensions perfectly, an effect on these variables is informative to understand the channels of the knowledge diffusion observed. Both outcomes would be indicative of strong connections with scientists within scientific communities. The two variables also differ substantially, as *collaboration*, and knowledge diffusion, in general, may occur from scientists that remain external to the firm, or, in the case of *Hiring*, via the actual mobility of scientists from academia (or other firms) to the focal firms. Table 7 presents the related results, from columns 1 to 3 for *Collaboration* and from columns to 4 to 5 for *Hiring*. We again include FE and control variables gradually as discussed for tables in section 7.

We find a strong and highly significant effect on *Collaboration*. The point estimates vary partly across specifications. In our preferred specifications, columns 2 and 3, the estimate implies, respectively, a 2.8 and 7.4 p.p higher probability of scientific collaborations between scientists and firms who participated in the same event. The magnitude of these effects is very large if compared with the variable conditional average of 4.8%. The correlation with the control variable remains as expected, as we see a higher probability of collaboration for scientists who have been previously cited by the firm or that are similar to the firm research. The magnitude of the effect of *Participation* relative to these correlations is meaningful. For instance, *Collaboration* is 15 p.p. higher for scientists having been cited by the firm. The magnitude of the effect of *Participation* is half of that.

We find no significant effect of *Participation* on *Hiring*. *Hiring* is also a rarer event, with a conditional average probability in the sample of 1.1%. However, the control variables show meaningful and significant correlations. For instance, indicating a much higher probability of mobility for scientists that have been previously cited by the firm.

These results allow us to conclude that actual collaborations with scientists are likely a relevant channel of knowledge diffusion. Also interestingly, the hiring of scientists is less likely to explain our results. It is interesting to note that this finding matches survey evidence from Cohen, Nelson, and Walsh (2002) who find that conferences and personal interactions score considerably higher than hiring in the importance as a channels of knowledge diffusion from public research to firms R&D. To be sure, by no means does this imply that hiring is not relevant in general. The mere fact that firms employ scientists, that we observe as authors of conference proceedings, implies that hiring plays a role. Further, at least sponsoring activities have among their stated objectives the promotion of job positions. Yet, mobility to firms could be a more geographically localized or long-term phenomenon, compared to knowledge diffusion and collaborations within international communities as we observe them at conferences. For most firms and scientists, participation in a conference may also not be a sufficiently strong treatment for a possibly more complex process as hiring. Moreover, the following sections demonstrate that for some firms, indeed an effect of *Participation* on *Hiring* exists. Finally, hiring of scientists to positions where they stop publishing may occur, which would not be observed in our data. However, the findings allow us to conclude that, on average, the observed effect on knowledge diffusion as well as on collaborations is explained by scientists that remain external to the firms.

Table 7: The effect of *Participation* on collaborations with and hiring of scientists

	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Hiring	(5) Hiring	(6) Hiring
Participation	0.032^{***} (0.009)	0.028^{***} (0.009)	0.074^{***} (0.020)	0.002 (0.004)	0.001 (0.004)	0.006 (0.008)
Science citations (L)			0.147^{***} (0.004)			0.037^{***} (0.001)
Patent citations (L)			0.074^{***} (0.005)			0.012^{***} (0.002)
Research similarity (L)			0.047^{***} (0.018)			0.021^{***} (0.007)
Conf. distance controls	No	No	Yes	No	No	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE		Yes	Yes		Yes	Yes
Year × Firm FE		Yes	Yes		Yes	Yes
R^2	0.177	0.185	0.196	0.075	0.080	0.087
Observations	5126376	5126273	5126273	5126376	5126273	5126273
Number clusters	1124	1114	1114	1124	1114	1114
DV cond. mean	0.048	0.048	0.048	0.011	0.011	0.011
F (First)	81.1	88.7	31.6	81.1	88.7	31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. In columns 1-3, the dependent variable is one if at least one of the proceeding authors has a joint publication with a firm researcher. In columns 4-6, the dependent variable is one if at least one of the proceeding authors becomes a firm researcher. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.

8.2 Firm participation intensity and research investments size

The effect of *Participation* likely depends on several margins related to firm characteristics and modes of participation in a conference. We focus here on the role of firms' participation intensity and firm size. Different levels of investments can be inferred from the number of proceedings presented and from sponsorship. We proxy firm size of investments in research by the number of active scientists in a year. This dimension has two main points of interest. First, from a theoretical standpoint and as discussed in section 2, we would expect different results if the prestige of firms within scientific communities were relevant, as opposed to the case where participation exclusively served the purpose of lowering knowledge search costs. Second, the role of this dimension has direct implications for firm decisions as well as for the resulting pattern of knowledge diffusion from science to industry.

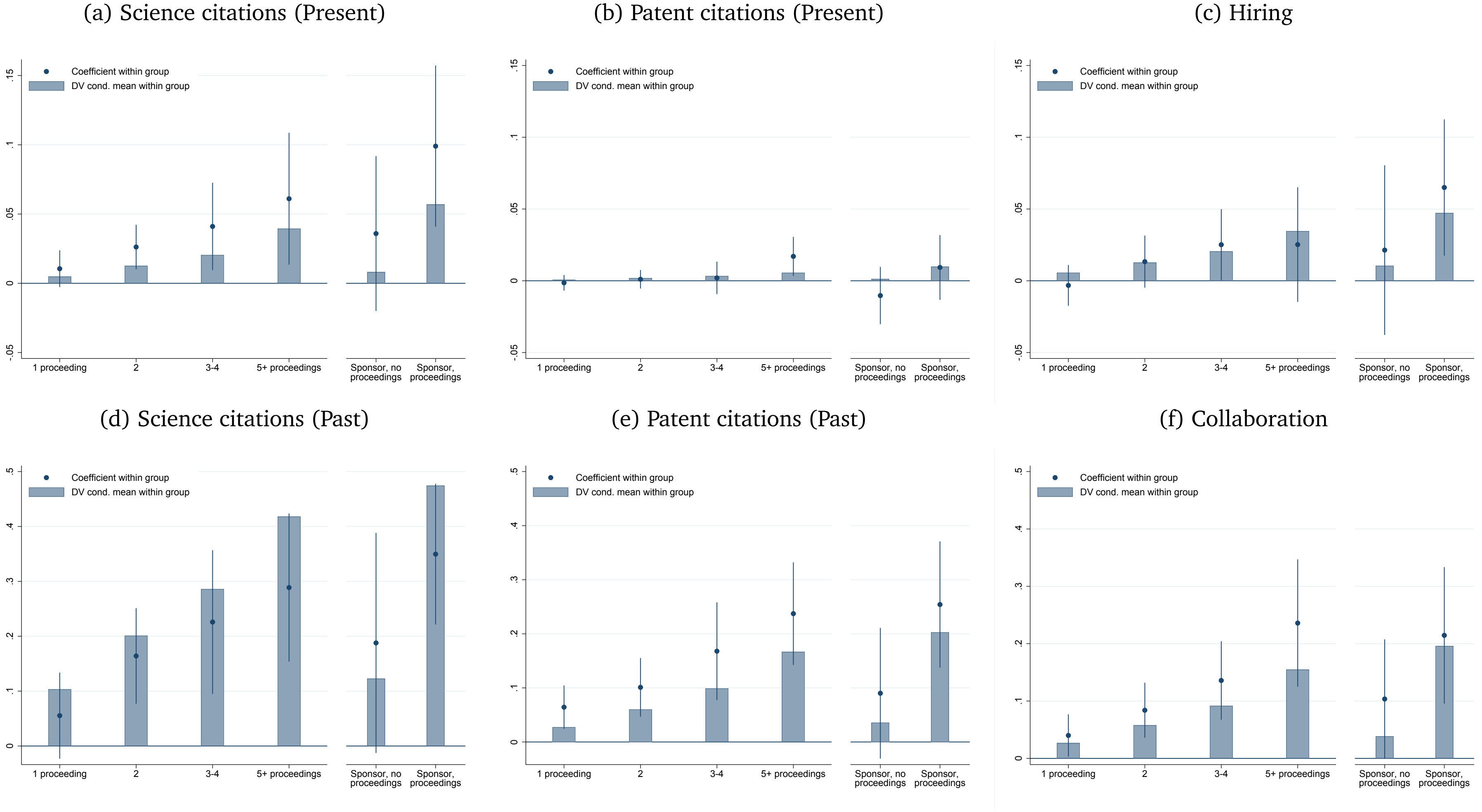
We study how the effect of *Participation* varies for firms that do not sponsor the focal conference and that have numbers of proceedings presented from low (1) to high (up to 5 or more), for firms that exclusively sponsor the conference, and for firms that both sponsor the conference and author proceedings. The underlying regression analyses are reported in Appendix Table A-6. For each outcome variable so far considered, we estimate interaction models. Both the endogenous variable, *Participation*, and the instrument *direct flight*, in the first-stage regressions, are interacted with dummy variables for each subgroup. We also estimate simplified regression models, where we introduce interactions only for sponsorship (Appendix Table A-4), or using the number proceedings as a single linear interaction variable, rather than four dummies for each subgroup (Appendix Table A-5). The results are a subset of those discussed here for the more complete model and lead to the same conclusions.

The results of this analysis are presented graphically. Figure 4 shows the main coefficient estimates from Appendix Table A-6 from the full model with all controls, but similar results are found in less complete specifications. Each graph, from (a) to (f), represents results for a different outcome variable, reporting the coefficient estimates with bandwidths for the 95% confidence intervals. The significance of the coefficient can be judged based on the distance from 0. The bars in the background indicate the within-group average of the dependent variable. The coefficients and group means are reported from left to right, for the effect of *Participation* for firms with 1, 2, 3 or 4 and 5 or more proceedings presented, and not sponsoring the focal conference, for firms that exclusively sponsor the conference, and for firms that both sponsor the conference and author at least one proceeding.

We refrain from commenting on every single coefficient and highlight the most relevant patterns. First, for all variables, we encounter a remarkably stronger effect for firms with a higher number of proceeding presented. Firms that only present one proceeding and that are not sponsors show no significant coefficients, except for the effect on patent citations to previous publications of the authors *Patent citations (Past)* and for *Collaboration* in graph (e) and (f). Firms with the largest number of proceedings, on the contrary, show the largest coefficients, that in most cases are also statistically significantly different from the estimated coefficients for firms with only one or 2 proceedings. For these firms, even an effect on patent citations to proceedings at the conference *Patent citations (Present)* can be found, compared to the average null effect (see Table 5).

Second, sponsorship and authorship appear complementary. Firms that only sponsor a conference event show no significant coefficient. The point estimates tend to be larger in magnitude and less precise, relative to coefficients for firms with only a few conference proceeding and not sponsoring, but never pass the significance threshold. For almost all variables, except for *Patent citations (Present)* in graph (b), firms that both sponsor a conference event and author at least a conference proceeding demonstrate a significant effect, higher or equal in magnitude to those of firms with the highest number of proceedings but not sponsoring. The comparison to the within-groups means of the dependent variables show that firms with higher intensity of participation also have on average higher levels of the dependent variables. The effect sizes remain substantial also relative to the within-groups sample averages. In other words, firms with higher levels of participation also have higher average probabilities citations and collaborations, regardless of the participation of scientists to the same conferences where they are present. However, *Participation* further adds to these levels.

Figure 4: Heterogeneity of the effect of *Participation* by participation intensity of the firm



(a) Science citations (Present)

(b) Patent citations (Present)

(c) Hiring

(d) Science citations (Past)

(e) Patent citations (Past)

(f) Collaboration

Notes: Each panel shows the result from an IV regression. We instrument each subgroup of firm participation intensity interacted with researcher participation is by the interaction of the same firm participation intensity and direct flight availability. For comparison, the dependent variable means in the subgroup of firm participation intensity are shown. Full estimation results are available in table A-6. Control variables and FE are as in other heterogeneity regressions.

We also highlight the interesting result of the positive and significant effect on *Hiring* for firms that are both sponsors and present proceedings (graph c). As noted, sponsoring has indeed also hiring among its objectives. However, by contrast, sponsorship alone shows no impact. This strongly suggests that the active participation of scientists is complementary and necessary to make sponsoring hiring activities effective. This is in line with the idea that interactions among scientists are a vehicle for exchanging information, possibly, in this case, also about scientific positions within firms.[19] Moreover, scientists assign value to the possibility of publishing and doing cutting edge research (Sauermann and Cohen, 2010; Stern, 2004). Firms incapable of signaling such opportunities may not be attractive for scientists. For the scope of our analyses, with reference to the discussion in the previous section 8.1, we maintain that hiring is unlikely to be an underlying mechanism explaining our results for the majority of firms. Instead, it is indeed possible that it contributes to strengthening the effects observed for both sponsoring and authoring firms.

In Table 8 we present results for the last analysis of heterogeneity of the effects. We look at the variation of the effects of *Participation* by firm level investments in research, as measured by the number of active scientists in a year. We categorize the size variable in Top 5 firms, Top 50 and the remainder of the sample. These are firms that in a given year have the highest number of active scientists (so they are often different firms over time). In the regressions, the interactions are based on values of the previous year. We present results for all outcome variables and our full specification from column 1 to 6. The results are generally coherent with the evidence from the intensity of participation, indicating that the largest firms have the strongest effects. Smaller firms have significant effects for patent citations to previous publications of scientists, but not for other variables. The estimates sizes for Top 5 firms are multiples of the effect size for smaller firms.

To conclude, the stronger effect of *Participation* for firms capable of strong contributions is striking. This is partly compatible with the notion of absorptive capacity, asserting that internal investments in research are necessary for effective learning. It is empirically impossible to tell apart precisely the role of the intensity of participation from other firm characteristics.[20] However, we posit that this strongly supports the interpretation that actual intense participation to a scientific community is a driver for the effects we observe. Firms with the lowest level of participation intensity almost show results “as if” they had not participated at all in a conference, and the interpretation that a proportional drop in knowledge flows would be observed if firms with higher levels of participation did not participate at all in conferences, is plausible. It remains to be acknowledged that we estimate results conditional on participation. We do not observe, nor could we causally identify the effect of participation for firms that do not participate at all in conferences. However, it turns out

[19] This interpretation is in line with insights from our interviews. Firms´ scientists declared that they participate in conferences mostly “like any other scientist” and do not normally have the objective of hiring other scientists on the behalf of the firm. At the same time, when firms sponsor the conference, HR personnel is often supported by firms scientists. Several HR representatives declared that exclusively sponsoring conference events turned out ineffective for hiring purposes, as it is difficult to attract the attention of potential candidates without displaying specific competencies and engagement into the scientific community.

[20] More generally, interactions models estimate the heterogeneity of the causal effect of an exogenous variable, but cannot be themselves causally interpreted.

that a very large number of firms have participated in at least one conference and are therefore in our sample. Moreover, the comparison of firms with low and high levels of participation intensity is informative. This finding also allows to speculate that the effect of firms that only passively attend, which we cannot observe, would be low or null.

Table 8: Heterogeneity of the effect of *Participation* by firms' size of research investments

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
Top 5 × Participation	0.097^{***}	0.241^{***}	0.016^{**}	0.278^{***}	0.156^{***}	0.004
	(0.024)	(0.067)	(0.008)	(0.056)	(0.040)	(0.019)
Top 6-50 × Participation	0.023^{***}	0.151^{***}	0.000	0.115^{***}	0.088^{***}	0.005
	(0.008)	(0.047)	(0.003)	(0.029)	(0.023)	(0.008)
Remainder × Participation	0.010	0.050	0.000	0.052^{**}	0.047^{**}	0.006
	(0.007)	(0.039)	(0.003)	(0.022)	(0.020)	(0.007)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Proceeding-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.069	0.370	0.027	0.173	0.195	0.087
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	11.0	11.0	11.0	11.0	11.0	11.0

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Ranks are yearly firms ranks, ordered by the size of the scientific workforce that based on publication information can be attributed to a company, lagged by one year. Individual coefficients of the rank levels are collinear with fixed effects and omitted. Proceeding-level controls include variables on firm citations before the conference towards the authors' previous publications, whether the proceeding is authored by a firm author, whether the authors have previously participated to the conference and the similarity of previous firm publications with the focal conference publication. Standard fixed effects include conference, origin × field, origin × firm, year × origin and year × firm fixed effects.

9 Robustness

Alternative citation measures

In this section, we address a list of possible concerns regarding the robustness of our results. We start considering alternative criteria for counting scientific citations and presenting results on a measure of text similarity as an alternative indicator of diffusion. Table 9 presents the related results, where the alternative variables are used as dependent variables of our full model specification. Column 1 reports, for reference, the results for our main dependent variable *Science cit.(present)*.

In column 2, we show results for a model where the dependent variable is the probability of observing citations to proceedings presented to the conference but excluding all author-level self-citations. Firm self-citations were already excluded in the main dependent variable. However, citations from scientists that move to the firm, and that were not affiliated to the firm at the moment of the conference, can still occur. As we have shown that hiring is weekly affected by participation, we doubt that this is frequent. Accordingly, the effect size in column two is very similar to the effect size in column one, suggesting that the main results are not driven by author self-citations.

In column 3, we look at citations only from publications where a scientist exclusively affiliated with the firm appears as the first author. This responds to the concern that the effect we observe may be driven by other academic scientists, or by other firms, that by coauthoring with the focal firm introduce citations that would not be otherwise observed. Indeed, coauthored publications are relatively frequent, especially with academics. In these cases, we cannot single out the individual contribution of single scientists of a focal firm to the bibliography list. To obviate this concern, we look at first-authored publications, under the assumption that the first author is the main project leader of a research project (this is indeed the normal practice in CS, also for scientists affiliated to firms). While the effect size (and the dependent variable mean) are somewhat smaller, the coefficient remains positive, statistically significant and still relatively large in magnitude.

Text similarity analysis

Scientific citations may reflect strategic behavior or salience (Teplitskiy et al., 2020), for instance if scientists add citations to please other scientists, or simply because the exposure to a conference proceeding increases the probability that that proceeding is cited rather than another, but without actual influence on the content. We alleviate this concern by showing that besides the probability of citations, the material content of the research subsequently carried out by the firm changes as well. We do so using a measure of text similarity between the focal proceeding p and future proceedings of the firm, within the same field and in a time window of 3 years. We discuss in detail the construction of the similarity measure in appendix F.

Column 4, in Table 9, presents the results for the text-similarity measure over the entire 3-year period and for the mean similarity with firm proceedings. Alternative specifications presented in appendices, show variants where we distinguish the similarity for each year separately (Table A-17) and using the maximum instead of the mean (Table A-18). The results are broadly consistent across these variants. From column 4, we see that *Participation* has a positive significant effect on the text-similarity measure of about 0.03, against a sample average of the variable of 0.1. This reinforces our finding that participation of firms to the scientific community has real and relevant effects on the firms' scientific activities.

Falsification test: effect on predetermined variables

To test the plausibility of our identification assumption, we design an analysis analogous to a falsification test. If the participation at the same conference is, after instrumenting, truly exogenous, then within our models, *Participation* should have no effect in these tests, while non-well identified models should still have a high probability to find significant estimates. First, as dependent variable we use citations from firm publications published before the conference to proceedings authored by the authors of proceeding p. Similarly, the measure of text similarity introduced in the section above should not show an effect for similarity with conference proceedings of the firm previous to the conference. In other words, we estimate whether there is any correlation between *Participation* and measures that are predetermined at the time of the conference. These measures are analogous to the control variables we use in our models, but we used them here as dependent variables. For patents, we only consider scientific references introduced by patent documents published before the conference year.

Table 10 shows the results. We pair the estimations from our IV-model with OLS regressions. In particular, column 1, 3 and 5 show OLS estimates, while columns 2, 4 and 6 estimates from the IV-models. Column 1 and 2 present results for the falsification test based on scientific citations. Column 3 and 4, based on patent citations. Column 5 and 6, based on text similarity. Notably, we can still find statistically significant coefficients, indicating endogeneity issues, in the OLS models. On the contrary, we do not encounter any statistically significant correlation in our IV-models. This increases our confidence that the instrumental variable strategy truly generates exogenous variation

Table 9: Robustness - Science citations' alternative measures

	(1) Science cit (present)	(2) Science cit (No self-cit)	(3) Science cit (First author)	(4) Similarity (Mean, Post)
Participation	0.021^{***}	0.018^{***}	0.014^{***}	0.030^{***}
	(0.007)	(0.007)	(0.005)	(0.007)
Science citations (L)	0.029^{***}	0.024^{***}	0.016^{***}	0.006^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Patent citations (L)	0.008^{***}	0.007^{***}	0.004^{***}	0.004^{***}
	(0.002)	(0.002)	(0.001)	(0.001)
Research similarity (L)	0.025^{***}	0.023^{***}	0.011^{***}	0.727^{***}
	(0.006)	(0.006)	(0.004)	(0.008)
Conf. distance controls	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes
R^2	0.083	0.076	0.054	0.768
Observations	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114
DV cond. mean	0.010	0.009	0.006	0.126
F (First)	31.6	31.6	31.6	31.5

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Column 4 also contains a control variable for the number of publications the similarity is computed for. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.

in the variable of interest and allows us to estimate causal effects.

A related insight is provided by the event study analysis in Appendix A.1, already mentioned in section 7.1. If either airlines' or conference organizers' decisions were driven endogenously by specific pair-level increases in participation from certain regions, this would likely emerge as a pre-trend in the probability of participation in the years prior to the change in availability of direct flights. We do not find any support for this concern.

Table 10: Falsification test - citations and similarity *before* the conference

	(1) Science cit (pseudo) OLS	(2) Science cit (pseudo) IV	(3) Patent cit (pseudo) OLS	(4) Patent cit (pseudo) IV	(5) Similarity (Mean, t-1) OLS	(6) Similarity (Mean, t-1) IV
Participation	0.029^{***}	0.022	0.002^{***}	0.010	0.016^{***}	−0.008
	(0.002)	(0.031)	(0.000)	(0.007)	(0.000)	(0.011)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.211	0.211	0.048	0.046	0.594	0.580
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.066	0.066	0.005	0.005	0.087	0.087
F (First)		28.9		28.9		28.5

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.Similarity refers to research similarity, measured by the similarity of abstracts of firm publications in the same subfield with the abstract of the focal proceeding.

Alternative model specifications

Finally, this section discusses additional results showing the robustness to alternative specifications, related to the inclusion of additional and different FE controls, and clustering of errors. The set of FE we include in our preferred specifications is meant to address the most reasonable concerns for our identification strategy. In Appendix, from Table A-7 to Table A-9 we show the robustness of our results to additional FE controls. Since conference locations are determined by unobserved amenities, the results may be driven by such factors. For example, it is evident how frequent conferences in Hawaii, Mallorca or in Florida are compared to the resident scientist population (Figure 1). Therefore, we show that the results are robust to the inclusion of conference location times year fixed effects (Table A-7). Consequently, the results are not driven by unobserved conference location amenities. Table A-8 shows robustness to the inclusion of origin × conference series fixed effects. This pair-level FE controls address possible concerns about the relative specificity of some conference series for scientists of certain regions. For instance, some conferences may have more national focus. This may influence the probability of participation of scientists regardless of accessibility and have implications also for the consequent interactions between scientists and firms. However, the inclusion of these FE controls does no change the results.

In table A-9, we control for the full interaction of firm, origin and year FE. This set of fixed effects captures every possible variation of connectivity and interactions between firms and particular regions as would be embodied for example in direct-flight connections between firms and the researcher regions, over time. It emphasizes the distinction between direct flights to conference locations (which we study) as opposed to firm regions (which we control for), and excludes any possibility that contemporaneous shocks between the firm and the scientists' regions pairs affect the results. In this model, the residual variation comes from the possibility that firms may be differently exposed to scientists within one same region (for instance in different fields of specialization), due to the different accessibility of these scientists to multiple conferences where the firms participate. Despite this specification being highly demanding, the results remain largely unaffected.

In table A-10, we test the robustness to various other cluster levels. Given the nature of the dataset, arguments can be made to cluster on the firm level (the acting entity in the second stage), the researcher location × conference location level (the level of the instrument) or the proceedings. We show that the standard errors in the second stage are not strongly affected by the cluster choice. However, the first stage often becomes substantially *more* powerful with alternative clusters. We have also tested whether the results are robust to using more control conferences (in particular, five instead of two) or to using a log specification instead of a linear probability model. Results are available from the authors upon request.

10 Conclusion

Participation of firms to scientific conferences, perhaps today at unprecedented levels of intensity in fields like Machine Learning (ML), is not a rare phenomenon. Over 20 years, from 1996 to 2010 in our analysis sample (and up to 2015 for the entire initial sample), we find constant and significant participation of firms, in terms of conference proceedings authored by firms scientists and sponsorship of conference events in the entire field of Computer Science (CS). The key finding of the paper is that participation decisions of firms and scientists to different conferences have a causal effect on firms' innovation outcomes. In other words, firms scientific and technological outcomes rely strongly on knowledge within the scientific communities in which they participate. The effect is not confined to citations to conference proceedings at the conference, but, especially for patents, extends to citations to previous work of scientists at the conference. We also find a strong effect on the probability of scientific collaborations, but not, on average, on hiring, which suggests that actual collaborations with external scientists are one important mechanism explaining our results. These effects are much stronger and significant for firms capable of highly intense participation, as captured by their sponsorship of conference events, the number of conference proceedings presented and their research investments size. Participation of a scientist to a conference where a firm is present both as sponsor and as author proceedings' affiliation also leads to higher probability of hiring. The effect for firms that make only minimal investments in participation is seldom significant.

The effect of active participation implies that physical proximity maintains an important role in the exchange of knowledge and the formation of collaborations. However, from a theoretical standpoint, the rationale for participation appears not confined to the need to abate search costs. Our finding is compatible with the theory that knowledge diffusion within science is shaped by its social norms and structure. The prestige of organizations within communities likely becomes a complementary asset that enables effective knowledge access, beyond the mere role of proximity. This provides the rationale for institutional investments in intense participation in scientific communities. This interpretation of the results is also supported by descriptive evidence and qualitative accounts on the nature of firms participation investments, and on their efforts to ensure that their contributions at scientific conferences are of the highest quality. The role of prestige appears all the more plausible if actual interactions and collaborations with external scientists, rather than the access to proceedings' content, is the main mechanism of diffusion.

These findings make various contributions. Similarly to previous studies, scientific communities appear to transcend organization boundaries, enabling knowledge flows between academia and industry and from science to technology. Also, the results confirm, on a larger scale and for knowledge flows across institutional boundaries, that scientists are strongly influenced by face-to-face interactions. The paper further suggests that active participation and prestige within the scientific community is an important antecedent to both collaborations and of knowledge flows, and may constitute a strategic objective for firms. Absorptive capacity appears not exclusively a function of cognitive ability or face-to-face interactions, but also of the prestige within scientific communities. Consequently, knowledge diffusion appears channeled towards firms with the highest level of participation.

Implications are substantial. First, the results offer a different perspective on the apparent paradox that investments of firms in research decrease while the relevance of science for firms remain stable, if not higher. Academia and industry interactions emerge to be significant and not unidirectional. Scientific communities remain an important platform that generates opportunities for the diffusion of knowledge. As such, the active participation of industry maintains an important role in the diffusion of knowledge from science to technology. The participation in scientific communities may increasingly be a way to access external knowledge, without internalizing fully its production. The results also challenge the notion that proximity and participation naturally lead knowledge spillovers to spread equally and freely. Institutions capable of intense participation are more likely to absorb knowledge, which in turn reinforces their ability to establish a position within the scientific community, in a process akin to the Matthew effect usually attributed to the accumulation of prestige of scientists or academic institutions. As a consequence, contributions of firms to scientific communities are not necessarily in conflict with firm objectives and may lead to the concentration of innovation capacity. At the same time, this may not be a viable strategy for most firms. Those that are only able of limited and short-lived investments may gain no benefit from interactions with scientific communities, and more limited returns from investing in research in general.

Finally, we highlight some limitations and directions for future research. First, the implications for firms' performance and general welfare implications are delicate. Firms capable of large investments in participation may benefit directly from scientific communities and at the same time use the opportunity to guide scientific advancements towards economically valuable applications. On the basis of the evidence on the value of science for firm value and the high value of science-based patents, we posit that connections to scientific communities likely bring great values to such firms. However, strong connections between firms and specific communities may have implications for the direction of research and the diversity of innovation options for firms, with not obvious long-run effects. Moreover, the evidence of concentrated knowledge flows may have implications for the competitive structure of science-based and high-tech industries, potentially increasing inequality. The scope of these considerations is limited by the microeconomic nature of our study and is left to future research. Second, the results suggest that knowledge flows across organizational boundaries are shaped by factors like group identity, reputation and prestige on social interactions that we cannot fully capture at our level of analyses. Related evidence at the individual level exists (Charness, Rigotti, and Rustichini, 2007; Chen and Li, 2009; Haeussler, 2011) which can be however extended to consider the interplay between the individual, organizational and institutional dimension. Finally, our study is limited to CS, although with data that are largely representative for this entire sector. Due to its importance for the economy, we tend to allege that evidence in this context is relevant. By virtue of descriptive evidence on the number of firms participating in conferences, and accounts regarding the relevance of science and conferences (e.g. in chemistry, pharmaceutics, biotechnology, engineering) (Cohen, Nelson, and Walsh, 2002), we also would expect that similar evidence would be found in other sectors. Extending this analysis to other contexts would still be recommended.

References

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein (2008). "Academic freedom, private-sector focus, and the process of innovation." In: *RAND Journal of Economics* 39.3, pp. 617–635 (cit. on p. 2).

Ahmadpoor, Mohammad and Benjamin F. Jones (2017). "The dual frontier: Patented inventions and prior scientific advance." In: *Science* 357.6351, pp. 583–587 (cit. on pp. 8, 19).

Akcigit, Ufuk, Santiago Caicedo Soler, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi (2018). "Dancing with the Stars: Innovation Through Interactions." In: *National Bureau of Economic Research* (cit. on p. 6).

Alcacer, Juan and Michelle Gittelman (2006). "Patent citations as a measure of knowledge flows: The influence of examiner citations." In: *Review of Economics and Statistics* 88.4, pp. 774–779 (cit. on p. 19).

Almeida, Paul and Bruce Kogut (1999). "Localization of Knowledge and the Mobility of Engineers in Regional Networks." In: *Management Science* 45.7, pp. 905–917 (cit. on p. 4).

Andrews, Dan, Chiara Criscuolo, and Peter N Gal (2015). “Frontier Firms, Technology Diffusion and Public Policy: Micro Evidence from OECD Countries.” In: *Working Paper*, p. 39 (cit. on p. 6).

Arora, Ashish, Sharon Belenzon, and Andrea Patacconi (2018). “The decline of science in corporate R&D.” In: *Strategic Management Journal* 39.1, pp. 3–32 (cit. on p. 5).

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2020). “Knowledge spillovers and corporate investment in scientific research.” In: *Working paper* (cit. on pp. 2, 5).

Arora, Ashish and Alfonso Gambardella (1994). “Evaluating technological information and utilizing it. Scientific knowledge, technological capability, and external linkages in biotechnology.” In: *Journal of Economic Behavior and Organization* 24.1, pp. 91–114 (cit. on pp. 2, 7).

Audretsch, David B. and Maryann P. Feldman (1996). “R&D Spillovers and the Geography of Innovation and Production.” In: *American Economic Review* 86.3, pp. 630–640 (cit. on p. 2).

Audretsch, David B. and Paula E. Stephan (1996). “Company-Scientist Locational Links : The Case of Biotechnology.” In: *American Economic Review* 86.3, pp. 641–652 (cit. on pp. 2, 4, 5).

Autor, David, David Dorn, Gordon H. Hanson, Gary Pisano, and Pian Shu (2020). “Foreign Competition and Domestic Innovation: Evidence from US Patents.” In: *American Economic Review: Insights*, Forthcoming (cit. on pp. 11, 71).

Bathelt, Harald and P. Cohendet (2014). “The creation of knowledge: local building, global accessing and economic development–toward an agenda.” In: *Journal of Economic Geography* 14.5, pp. 869–882 (cit. on p. 6).

Belenzon, Sharon and Mark Schankerman (2013). “Spreading the Word: Geography, Policy, and Knowledge Spillovers.” en. In: *Review of Economics and Statistics* 95.3, pp. 884–903 (cit. on p. 2).

Bhattacharya, Haimanti and Subhasish Dugar (2014). “Partnership formation: The role of social status.” In: *Management Science* 60.5, pp. 1130–1147 (cit. on p. 6).

Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti (2017). “Researcher’s Dilemma.” In: *Review of Economic Studies* 84.3, pp. 969–1014 (cit. on p. 6).

Boudreau, Kevin J., Tom Brady, Ina Ganguli, Patrick Gaule, Eva Guinan, Anthony Hollenberg, and Karim R. Lakhani (2017). “A field experiment on search costs and the formation of scientific collaborations.” In: *Review of Economics and Statistics* 99.4, pp. 565–576 (cit. on pp. 2, 5, 6).

Breschi, Stefano and Francesco Lissoni (2001). “Knowledge spillovers and local innovation systems: a critical survey.” In: *Industrial and Corporate Change* 10.4, p. 975 (cit. on p. 5).

Brynjolfsson, Erik and Lorin M. Hitt (2003). “Computing productivity: Firm-level evidence.” In: *Review of Economics and Statistics* 85.4, pp. 793–808 (cit. on pp. 3, 7).

Campante, Filipe and David Yanagizawa-Drott (2017). “Long-Range Growth: Economic Development in the Global Network of Air Links.” In: *The Quarterly Journal of Economics* (cit. on pp. 16, 18).

Campos, Raquel, Fernanda Leite Lopez de Leon, and Ben McQuillin (2018). “Lost in the Storm: The Academic Collaborations That Went Missing in Hurricane ISSAC.” In: *Economic Journal* 128.610, pp. 995–1018 (cit. on pp. 2, 5, 6).

Cassiman, Bruno and Reinhilde Veugelers (2002). “R&D cooperation and spillovers: some empirical evidence from Belgium.” In: *American Economic Review* 92.4, pp. 1169–1184 (cit. on pp. 2, 5, 7).

— (2006). “In Search of Complementarity in Innovation Strategy: Internal R&D and External Knowledge Acquisition.” In: *Management Science* 52.1, pp. 68–82 (cit. on pp. 2, 5, 7).

Catalini, Christian, Christian Fons-Rosen, and Patrick Gaulé (2020). “How Do Travel Costs Shape Collaboration?” In: *Management Science*, pp. 1–21 (cit. on pp. 3, 16).

Cavacini, Antonio (2015). “What is the best database for computer science journal articles?” In: *Scientometrics* 102.3, pp. 2059–2071 (cit. on pp. 67, 68).

Cetina, Karin Knorr (1999). *Epistemic cultures: how the sciences make knowledge*. Ed. by Harvard University Press. Vol. 53. 9. Harvard University Press, pp. 1689–1699 (cit. on pp. 2, 6).

Chai, Sen and Richard B. Freeman (2019). “Temporary colocation and collaborative discovery: Who confers at conferences.” In: *Strategic Management Journal* 40.13, pp. 2138–2164 (cit. on pp. 2, 5, 6).

Charness, Gary, Luca Rigotti, and Aldo Rustichini (2007). “Individual Behavior and Group Membership.” In: *American Economic Review* 97, pp. 1340–1352 (cit. on pp. 2, 6, 39).

Chen, Yan and Sherry Xin Li (2009). “Group identity and social preferences.” In: *American Economic Review* 99.1, pp. 431–457 (cit. on pp. 6, 39).

Cockburn, Iain M. and Rebecca Henderson (1998). “Absorptive capcity, coauthoring behavior, and the organization of research in drug discovery.” In: *The Journal of Industrial Economics* 46.2, pp. 157–182 (cit. on pp. 2, 4, 5, 7).

Cohen, Wesley M. and Daniel A. Levinthal (1989). “Innovation and Learning: The Two Faces of R&D.” In: *The Economic Journal* 99.397, p. 569 (cit. on pp. 2, 7).

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh (2002). “Links and impacts: The influence of public research on industrial R&D.” In: *Management Science* 48.1, pp. 1–23 (cit. on pp. 5, 8, 28, 39).

Crane, Diana (1974). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Vol. 15. 1, p. 139 (cit. on pp. 2, 6).

Dasgupta, Partha and Paul A. David (1994). “Toward a new economics of science.” In: *Research Policy* 23.5, pp. 487–521 (cit. on pp. 2, 4, 6).

Eeckhout, Jan and Boyan Jovanovic (2002). “Knowledge spillovers and inequality.” In: *American Economic Review* 92.5, pp. 1290–1307 (cit. on p. 5).

Fleming, Lee, H. Greene, G. Li, Matt Marx, and D. Yao (2019). “Government-funded research increasingly fuels innovation.” In: *Science* 364.6446, pp. 1139–1141 (cit. on p. 5).

Fleming, Lee and Olav Sorenson (2004). “Science as a map in technological search.” In: *Strategic Management Journal* 25.89, pp. 909–928 (cit. on p. 6).

Franceschet, Massimo (2010). “The role of conference publications in CS.” In: *Communications of the ACM* 53.12, pp. 129–132 (cit. on pp. 3, 8).

Giroud, Xavier (2013). “Proximity and investment: Evidence from plant-level data.” en. In: *Quarterly Journal of Economics* 128.2, pp. 861–915 (cit. on pp. 3, 16, 18).

Gittelman, Michelle (2007). “Does geography matter for science-based firms? Epistemic communities and the geography of research and patenting in biotechnology.” In: *Organization Science* 18.4, pp. 724–741 (cit. on pp. 2, 5, 7).

Gittelman, Michelle and Bruce Kogut (2003). “Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns.” In: *Management Science* 49.4, pp. 366–382 (cit. on pp. 2, 7).

Gregg, Forest and Derek Eder (2019). “Dedupe.” https://github.com/dedupeio/dedupe (cit. on p. 71).

Grossman, Gene M and Elhanan Helpman (1993). *Innovation and Growth in the Global Economy*. Vol. 1. The MIT Press (cit. on p. 2).

Haeussler, Carolin (2011). “Information-sharing in academia and the industry: A comparative study.” In: *Research Policy* 40.1, pp. 105–122 (cit. on pp. 2, 6, 39).

Haeussler, Carolin, Lin Jiang, Jerry Thursby, and Marie Thursby (2014). “Specific and general information sharing among competing academic researchers.” In: *Research Policy* 43.3, pp. 465–475 (cit. on pp. 2, 4, 6).

Hartmann, Philipp and Joachim Henkel (2020). “The Rise of Corporate Science in AI: Data as a Strategic Resource.” In: *Academy of Management Discoveries* (cit. on p. 8).

Hellmann, Thomas and Enrico Perotti (2011). “The circulation of ideas in firms and markets.” In: *Management Science* 57.10, pp. 1813–1826 (cit. on p. 7).

Hicks, Diana (1995). “Published papers, tacit competencies and corporate management of the public/private character of knowledge.” In: *Industrial and Corporate Change* 4.2, pp. 401–424 (cit. on pp. 2, 5, 6).

Jaffe, Adam B (1989). “Real effects of academic research.” In: *American Economic Review*, pp. 957–970 (cit. on pp. 2, 5).

Jaffe, Adam B, M. Trajtenberg, and Rebecca Henderson (1993). “Geographic localization of knowledge spillovers as evidenced by patent citations.” In: *Quarterly Journal of Economics* 108.3, pp. 577–598 (cit. on pp. 2, 6, 19).

Jones, Benjamin F. (2009). “The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?” In: *Review of Economic Studies* 76.1, pp. 283–317 (cit. on p. 6).

Kim, Jinseok (2018). “Evaluating author name disambiguation for digital libraries : a case of DBLP.” In: *Scientometrics* (cit. on p. 67).

Knaus, Johannes and Margit Palzenberger (2018). “PARMA. A full text search based method for matching non-patent literature citations with scientiic reference databases. A pilot study. Technical Report by the Max Planck Digital Library, Big Data Analytics Group.” In: (cit. on pp. 12, 67).

Laursen, Keld and Ammon J. Salter (2006). “Open for innovation: The role of openness in explaining innovation performance among U.K. manufacturing firms.” In: *Strategic Management Journal* 27.2, pp. 131–150 (cit. on p. 7).

Lopez de Leon, Fernanda Leite and Ben McQuillin (2020). “The Role of Conferences on the Pathway to Academic Impact.” In: *Journal of Human Resources* 55.1, pp. 164–193 (cit. on pp. 2, 5, 6).

Maskell, Peter (2014). “Accessing remote knowledge-the roles of trade fairs, pipelines, crowdsourcing and listening posts.” In: *Journal of Economic Geography* 14.5, pp. 883–902 (cit. on p. 6).

Merton, Robert K (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press (cit. on pp. 2, 6, 7).

Mukherjee, Arijit and Scott Stern (2009). “Disclosure or secrecy? The dynamics of Open Science.” In: *International Journal of Industrial Organization* 27.3, pp. 449–462 (cit. on p. 6).

Nelson, Richard R. (1962). “The Link Between Science and Invention: The Case of the Transistor.” In: *The Rate and Direction of Inventive Activity*. Vol. ISBN, pp. 549–584 (cit. on pp. 3, 8).

Newell, Allen, Alan J. Perlis, and Herbert A. Simon (1967). “Computer science.” In: *Science* 157.3795, pp. 1373–1374 (cit. on p. 7).

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi (2019). “Science quality and the value of inventions.” In: *Science Advances* 5.12, eaay7323 (cit. on pp. 12, 19, 67).

Roach, Michael and Wesley M. Cohen (2013). “Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research.” In: *Management Science* 59.2, pp. 504–525 (cit. on pp. 19, 20).

Romer, Paul M (1990). “Endogenous technological change.” In: *Journal of Political Economy* 98.5, S71–S102 (cit. on p. 2).

Rosenberg, Nathan (1990). “Why do firms do basic research (with their own money)?” In: *Research Policy* 19.2, pp. 165–174 (cit. on pp. 2, 5).

Rosenkopf, Lori and Atul Nerkar (2001). "Beyond local search: Boundary-spanning, exploration, and impact in the optical disk industry." In: *Strategic Management Journal* 22.4, pp. 287–306 (cit. on p. 6).

Sauermann, Henry and Wesley M. Cohen (2010). "What Makes Them Tick? Employee Motives and Firm Innovation." In: *Management Science* 56.12, pp. 2134–2153 (cit. on p. 32).

Schmidheiny, Kurt and Sebastian Siegloch (2020). "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization." In: *SSRN Electronic Journal* (cit. on p. 45).

Schnitzer, Monika and Martin Watzinger (2019). "Standing on the Shoulders of Science." In: *CEPR Discussion Paper* No. DP1376 (cit. on p. 19).

Simeth, Markus and Michele Cincera (2016). "Corporate Science, Innovation, and Firm Value." In: *Management Science* 62.7, pp. 1970–1981 (cit. on p. 5).

Simeth, Markus and Julio D. Raffo (2013). "What makes companies pursue an Open Science strategy?" In: *Research Policy* 42.9, pp. 1531–1543 (cit. on pp. 2, 5).

Stein, Jeremy C. (2008). "Conversations among competitors." In: *American Economic Review* 98.5, pp. 2150–2162 (cit. on pp. 5, 6).

Stephan, Paula E. (1996). "The economics of science." In: *Journal of Economic literature* 34.3, pp. 1199–1235 (cit. on pp. 2, 4, 6, 7).

Stern, Scott (2004). "Do Scientists Pay to Be Scientists?" In: *Management Science* 50.6, pp. 835–853 (cit. on p. 32).

Teplitskiy, Misha, Eamon Duede, Michael Menietti, and Karim R. Lakhani (2020). "Citations Systematically Misrepresent the Quality and Impact of Research Articles: Survey and Experimental Evidence from Thousands of Citers." In: (cit. on pp. 20, 34).

Thompson, Peter and Melanie Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." In: *American Economic Review* 95.1, pp. 450–460 (cit. on p. 19).

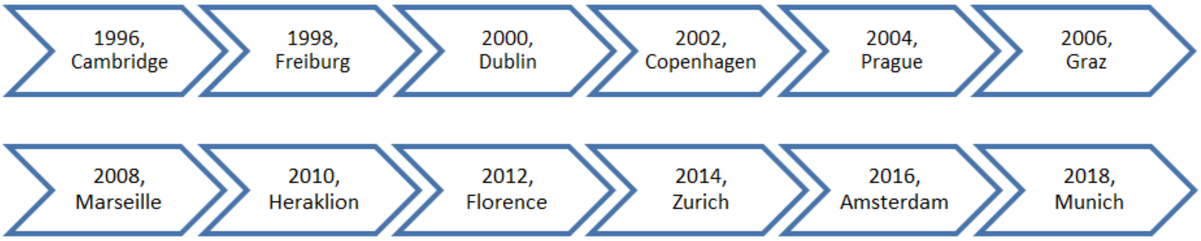
Vlasov, Stanislav A., Marc D. Bahlmann, and Joris Knoben (2016). "A study of how diversity in conference participation relates to SMEs' innovative performance." In: *Journal of Economic Geography*, lbw004 (cit. on p. 2).

A Appendix: Investigating the first stage

A.1 Event study on the effect of direct flights on participation

We design an event study on the dynamics of the effect of direct flights on the probability of participation of scientists to conferences. The variation in the availability of direct flights derives from new airlines routes and from the changes in venues of conference series. As an example, figure 5 shows the case of the European Conference on Computer Vision (ECCV), an *A*-ranked conference in AI/Machine vision. The conference takes place every two years at varying places in Europe and the locations are decided roughly with 4 years of advance. In our most conservative models, the identifying variations would derive from the possibility to access the conference venues with direct flights, for scientists outside the country where the conference is hosted, after controlling for time-specific FE of their locations, and for their geographic distance. In robustness analyses we also control for the specific conference events fixed effects.

In the event study setting we can explore, for instance, whether any anticipation in participation exists. We collapse our data and build a panel at the level of scientists locations and conference series pair-level. We use as dependent variables the number of researchers from a region that participate to the conference in a given time period and, alternatively, a dummy equal 1 if at least one scientists from that region participates. We construct variables on the change of direct flight availability. If in this period, relative to the previous period, a direct flight connection to the conference series becomes available, the direct flight indicator is 1. If a direct flight connection is no longer available, the indicator is -1. If there is no change, the indicator is 0. Note that some conferences like the ECCV do not occur every year, which is why we are using relative time periods.



Notes: ECCV: European Conference on Computer Vision. We visited this conference in 2018 and discuss findings in section 3.

Figure 5: ECCV locations over time

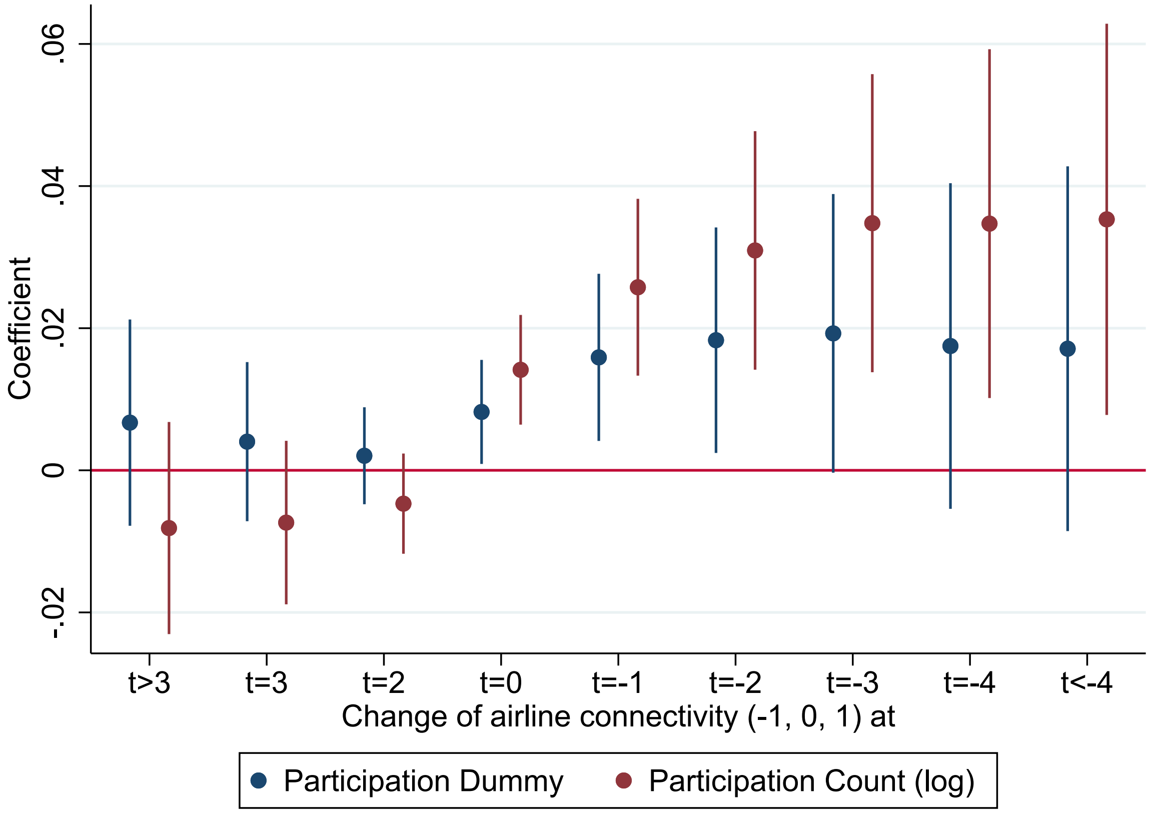
$$y_{rct} = \sum_{j=-4}^{5} \gamma_j D^j_{rct} + \mu_{rt} + \mu_{rc} + \mu_{ct} + \beta X_{rct} + \epsilon_{rct} \qquad \text{(A.1)}$$

More formally, we look at a panel of researchers' region r, conference series c by time period t in an event-study setting. The endpoints are binned, following the suggestion of Schmidheiny and Siegloch (2020). For each period j, the variable D^j_{rct} takes value 1 if a direct flight is introduced, value -1 if a direct flight is removed, and 0 if no change occurs. The coefficient γ_j captures the effect of a positive change. The period $j-1$ is used as baseline. In addition, we control for fixed effects on the region-year, region-conference series and conference series-year (conference event) level. We add the same control variables on a researcher region-conference event-level which are also employed in the main text. We cluster on the conference series-researcher region level.

We restrict the combinations to all region-conference series combinations to various sets of candidate regions. The most narrow set is defined by all regions from which at least one researcher ever attends the focal conference. The results from this specification are shown in figure 6 and columns

3/6 of table 11. The second definition considers all location with at least one researcher active in the conference's field. This is shown in columns 2/5. Finally, the third definition considers all possible locations. Here, columns 1/4 of table 11 are relevant.

Figure 6 plots the regression results of equation A.1, for log-transformed participation counts and the participation binary dummy variable. Table 11 lists detailed estimation results for the log-transformed counts. Results for the dummy variable as dependent variable are equivalent and available upon request. There is a relevant and statistically significant increase of the number of participants. This effect starts immediately at the point of the flight introduction and remains present, but no pre-trends can be found. The flight-induced participation seems to be persistent and slightly increasing in subsequent years if a direct flight connection persists. Based on the regression results, pre-trends or anticipation effects are likely only not a concern. All other coefficients behave as expected. When the conference is in their home region, researchers are more likely to attend. Researchers are less likely to attend distant conferences.



Notes: Coefficients from linear regression with 95% confidence intervals. Estimation for years 1996-2015. FE for region×year, region× conference series and conference events included. Other controls include distance (log), indicator variables for conference being held at the researcher region and for domestic flight connections. Clustering is on region×conference series level.

Figure 6: Event study estimates of the effect of direct flights on participation (Region-conference series level of analysis)

Table 11: Event study regression results of the effect of direct flights on participation (Region-conference series level of analysis)

	(1)	(2)	(3)	(4)	(5)	(6)
Participants	log(1+)	log(1+)	log(1+)	log(1+)	log(1+)	log(1+)
Domestic (State)	0.012***	0.018***	0.076***	0.012***	0.019***	0.082***
	(0.001)	(0.001)	(0.006)	(0.001)	(0.002)	(0.007)
Same region	0.169***	0.159***	0.044**	0.183***	0.173***	0.048**
	(0.014)	(0.014)	(0.019)	(0.017)	(0.017)	(0.023)
Distance (log)	−0.002***	−0.003***	−0.018***	−0.002***	−0.003***	−0.019***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)
t>3				−0.001	−0.001	−0.008
				(0.002)	(0.003)	(0.008)
t>2	−0.001	−0.001	0.000			
	(0.001)	(0.001)	(0.004)			
t=3				−0.002	−0.003	−0.007
				(0.002)	(0.002)	(0.006)
t=2	−0.001	−0.001	0.000	−0.002	−0.002*	−0.005
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.004)
t=0	0.004***	0.005***	0.010***	0.005***	0.006***	0.014***
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.004)
t=-1	0.007***	0.009***	0.017***	0.010***	0.012***	0.026***
	(0.001)	(0.002)	(0.005)	(0.002)	(0.002)	(0.006)
t=-2	0.007***	0.009***	0.018***	0.011***	0.012***	0.031***
	(0.002)	(0.002)	(0.006)	(0.002)	(0.003)	(0.009)
t=-3	0.007***	0.008***	0.018**	0.011***	0.013***	0.035***
	(0.002)	(0.002)	(0.007)	(0.003)	(0.004)	(0.011)
t=-4				0.011***	0.012***	0.035***
				(0.003)	(0.004)	(0.013)
t<-3	0.006***	0.007**	0.014*			
	(0.002)	(0.003)	(0.008)			
t<-4				0.009**	0.010**	0.035**
				(0.004)	(0.005)	(0.014)
Region set	All	Field	Attendance	All	Field	Attendance
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Conf. Ser. FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Conf. Ser. FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.467	0.473	0.505	0.519	0.523	0.534
Observations	4014488	2318856	303520	2600869	1491435	202066
Number clusters	668026	390677	48110	485981	281388	36601

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parentheses, clustered at the conference series times researcher region level. Regression results for equation A.1. In columns 3/6 all regions from which at least one researcher ever attends the focal conference are considered. All location with at least one researcher active in the conference's field are considered in columns 2/5. All possible locations are considered in columns 1/4.

A.2 Appendix: First stage - heterogeneity of the effects of *Direct Flight* on *Participation*

We investigate how the strength of the instrumental variable varies in the first stage by interacting it with key characteristics of conferences. We estimate regression A.2 with various sets of heterogeneity dimensions h. Generally speaking, we expect direct flights to matter more when the researchers are otherwise indifferent between similar conferences along our matching criteria. For example, we expect researchers to spare less to attend $A^{\star}$ or A conferences. For lower-level conferences, the discomfort of traveling might start to play a stronger role. Consequently, we expect the instrument to have less relevance for highly-ranked conferences and more for low-ranked conferences. Similarly, the availability of direct flights should matter more at long distances. We maintain the same level of observation then for our main analyses in the paper, detailed in section 6.

$$\begin{aligned} D\{\text{p presented at c}\}_{fpc} = & \sum \beta_{1i} D\{\text{direct flight}\}_{fpc} \times h_{pi} \\ & + \sum \beta_{2i} log\,(1+\text{distance})_{fpc} \times h_{pi} \\ & + \sum \beta_{2i} D\{\text{direct flight}\}_{fpc} \times log\,(1+\text{distance})_{fpc} \times h_{pi} \\ & + \sum \beta_{3i} h_{pi} + \beta_4 X_{fpc} + u_{fpc} \end{aligned} \tag{A.2}$$

Results are presented in Table 12. Column 1 corresponds to our main specification. Column 2 shows that the effect of the instrument is larger for conferences at longer distances. Also as expected, it turns out that quality level of the conference matters. For models without distance controls (column 3), the instrumental variable is significant for all rank levels of conferences, but it is already stronger in magnitude for lower ranked-conferences. In the most conservative models, including distance controls, the coefficient size for $A^{\star}$ and A conferences is comparatively small and weekly significant (column 4). However, the strength of the instrument increases for longer geographic distances and becomes large and significant at long distances for all quality levels. We see this in column 5 where we interact the distance of conferences with the effect of *Direct flight* for conferences of different ranking (triple interactions). In these models, the value of the distance variable is centered at the mean, so that the coefficient on the interacting variables can be interpreted as effects at the mean (corresponding to approximately 3600 km, 8.2 in logarithm).

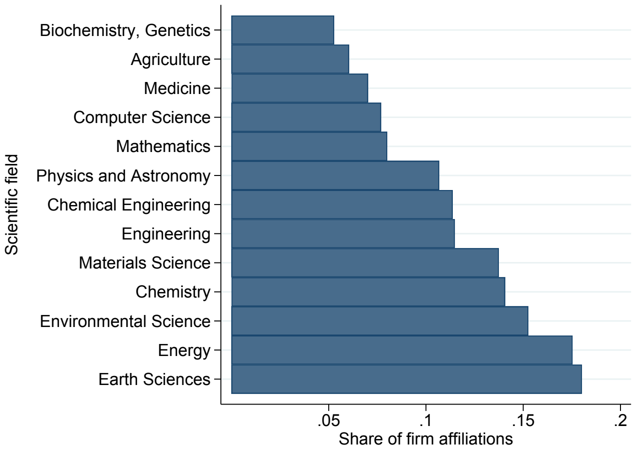
Table 12: First stage - Heterogeneity of the effect of *Direct flight* on *Participation*

	(1) Participation	(2) Participation	(3) Participation	(4) Participation	(5) Participation
Direct flight	0.030***	0.033***			
	(0.006)	(0.006)			
$A^{\star}$/A-level × DF			0.039***	0.010*	0.018***
			(0.005)	(0.005)	(0.005)
B/C-level × DF			0.077***	0.049***	0.046***
			(0.009)	(0.008)	(0.008)
log(Distance)	−0.039***	−0.048***		−0.039***	
	(0.003)	(0.004)		(0.003)	
$A^{\star}$/A-level × log(Distance)					−0.039***
					(0.004)
B/C-level × log(Distance)					−0.055***
					(0.005)
Direct flight × log(Distance)		0.023***			
		(0.005)			
$A^{\star}$/A-level × DF × log(d)					0.024***
					(0.006)
B/C-level × DF × log(d)					0.020***
					(0.005)
Same airport	−0.164***	−0.231***		−0.164***	−0.235***
	(0.036)	(0.039)		(0.036)	(0.039)
Domestic (State)	0.132***	0.128***		0.132***	0.126***
	(0.015)	(0.015)		(0.015)	(0.015)
Conf Ser FE	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes
Cluster	Origin	Origin	Origin	Origin	Origin
Number clusters	1114	1114	1114	1114	1114
R^2	0.333	0.334	0.323	0.334	0.334
Observations	5126273	5126273	5126273	5126273	5126273

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. The dataset follows the description of table 4. The value of log(Distance) is centered at the mean value in the regressions. The sample mean of log(Distance) is about 8.2, corresponding to approximately 3600 km.

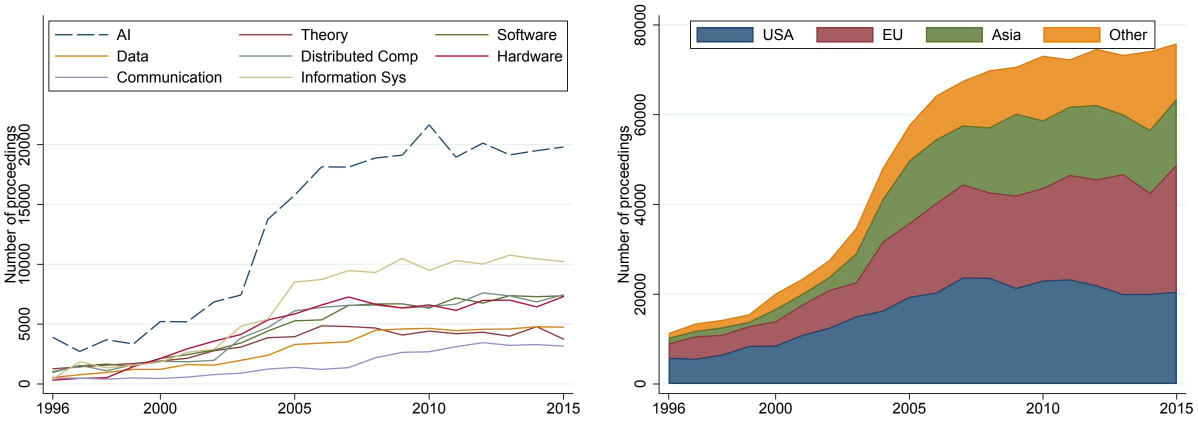
B Appendix: Figures

Figure A-1: Share of firms-authored conference proceedings by field



Notes: Fields are identified based on ASJC codes from Scopus between 1996-2015. Largest fields (millions of proceedings) are Engineering (2.19), CS (1.49), Physics (0.53), Mathematics (0.30), Material Science (0.28), Energy (0.18). Smallest fields are Agriculture (0.02), Biochemistry (0.02), Chemistry (0.03), Medicine (0.03), Environmental Science (0.08). Fields with less than 15,000 items or in social sciences or humanities are disregarded.

Figure A-2: Conference proceedings by field and conference locations



(a) Size of different CS fields

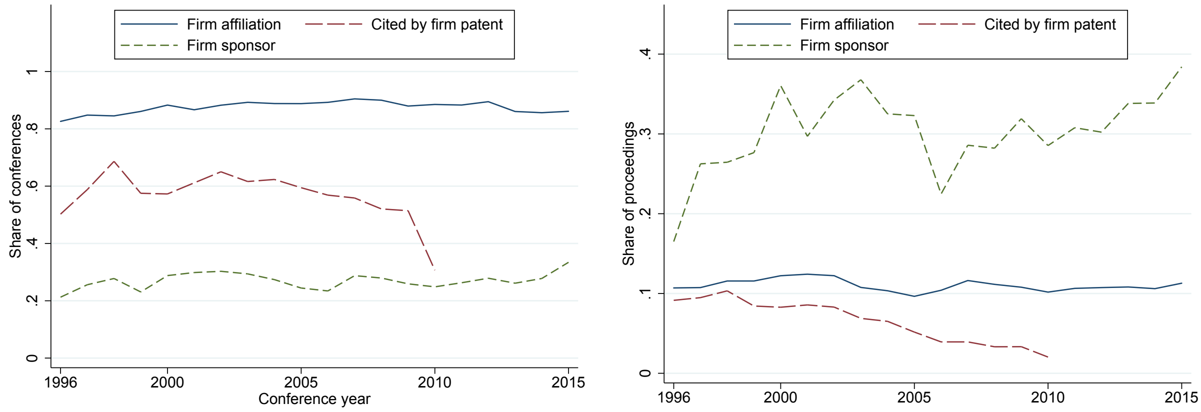
(b) Conference locations

Notes: Only DBLP conferences with matched Web of Science/Scopus articles as well as CORE information are considered. Conferences in panel A-2a are assigned to their first CORE field code.

Figure A-3: Firm participation and patent citations

(a) Share of conferences

(b) Share of proceedings

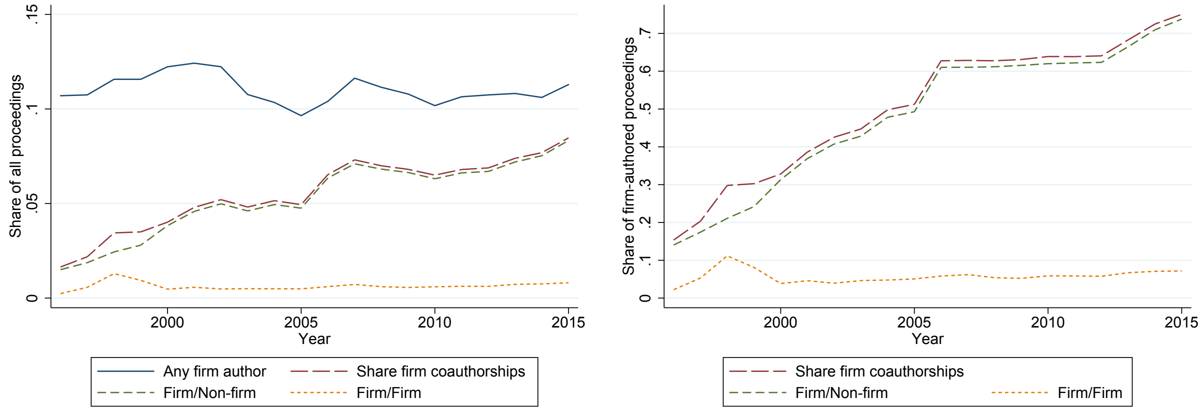


Notes: Shows activity of firms on the conference (A-3a) and proceeding (A-3b) level. In A-3b, sponsorship refers to a proceeding at a conference with at least one firm sponsor. Due to truncation we restrict the data based on patent citations to the year 2010 and before. The negative tendency is still likely the artifact of citations data truncation.

Figure A-4: Coauthorships of CS proceedings with a firm

(a) Coauthorships with a firm overall

(b) Coauthorships on firm proceedings



Notes: Coauthorships of firms a different firm or institution on conference proceedings. In A-4a, the share of all proceedings with firm authors is decomposed in proceedings with coauthorships and such without, further split up in such with firm-firm and firm-academia coauthorships. A-4b plots the same decomposition conditional on having at least one firm author.

C Appendix Tables

Table A-1: Covariate balancing table.

	Average value	Difference (SE)		p-value
	(Actual conf)	(Counterfactual-Actual)		
Exact matching criteria				
Year	2005.22	0.000	(0.000)	1.00
Rank: A*	0.10	0.000	(0.000)	1.00
Rank: A	0.28	0.000	(0.000)	1.00
Rank: B	0.36	0.000	(0.000)	1.00
Rank: C	0.25	0.000	(0.000)	1.00
Field: General CompSci	0.08	0.000	(0.003)	0.94
Field: General Engineering	0.05	−0.001	(0.001)	0.46
Field: AI / Computer Vision	0.21	0.000	(0.001)	1.00
Field: Computation Theory	0.16	0.000	(0.001)	1.00
Field: Computer Software	0.22	0.000	(0.001)	1.00
Field: Data Format	0.10	−0.001	(0.002)	0.55
Field: Distributed Computing	0.13	0.000	(0.001)	0.80
Field: Information Systems	0.13	−0.002	(0.002)	0.49
Coarsened matching criteria				
Size of the conference	70.40	0.022	(0.563)	0.97
Mean 5-year citations	4.46	0.005	(0.056)	0.93
Untargeted matching criteria				
Conference series age	5.27	−0.042	(0.052)	0.42
Number of fields	1.10	−0.002	(0.006)	0.70
Number of sponsors	1.34	−0.010	(0.018)	0.56
Number of firms	4.95	−0.021	(0.059)	0.72
Observations	5799	10492		

Notes: Covariate balancing for two counterfactual conferences. Shows the average deviation of the counterfactual conference from the actual conference.

Table A-2: CORE fields.

	Share (first)	Count	Share (freq)	Count
Computer Science (general)	7.3	41754	7.1	40476
Engineering (general)	15.1	85837	15.4	87747
Design (general)	0.0	0	0.1	518
Artificial Intelligence and Image Processing	28.8	163843	28.7	163258
Computation Theory and Mathematics	8.2	46612	8.1	46213
Computer Software	10.2	57895	10.1	57468
Data Format	6.4	36332	6.4	36427
Distributed Computing	10.2	58017	9.6	54782
Information Systems	13.8	78812	13.9	79252
Library and Information Studies	0.0	0	0.5	2962
Total	100.0	569102	100.0	569102

Notes: CORE fields as aggregated in the conference-level match are shown. Each conference series is associated with up to three CORE fields. Shares and counts using the first or using equal weighting among the CORE fields is shown. 1996-2015 data is shown.

Table A-3: Scientific and commercial value of corporate proceedings.

log 5y	(1) Science Citations	(2) Science Citations	(3) Science Citations	(4) Science Citations	(5) Science Citations	(6) Science Citations
Firm	0.262^{***}	0.239^{***}	0.063^{***}	0.056^{***}	0.050^{***}	0.044^{***}
	(0.011)	(0.011)	(0.006)	(0.006)	(0.005)	(0.005)
Sponsor	0.088^{***}	0.079^{***}	0.008	0.005		
	(0.023)	(0.023)	(0.014)	(0.014)		
Firm=Sponsor		0.389^{***}		0.125^{***}		0.109^{***}
		(0.044)		(0.028)		(0.024)
Year FE	Yes	Yes	Yes	Yes		
Conf FE					Yes	Yes
Conf Series FE			Yes	Yes		
R^2	0.015	0.016	0.260	0.260	0.326	0.326
Clusters	7298	7298	7295	7295	7217	7217
Obs	612103	612103	612100	612100	612022	612022

log 5y	(1) Patent Citations	(2) Patent Citations	(3) Patent Citations	(4) Patent Citations	(5) Patent Citations	(6) Patent Citations
Firm	0.095^{***}	0.091^{***}	0.069^{***}	0.066^{***}	0.067^{***}	0.064^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Sponsor	0.018^{***}	0.016^{***}	0.012^{***}	0.011^{***}		
	(0.004)	(0.004)	(0.003)	(0.003)		
Firm=Sponsor		0.072^{***}		0.055^{***}		0.038^{***}
		(0.014)		(0.010)		(0.009)
Year FE	Yes	Yes	Yes	Yes		
Conf FE					Yes	Yes
Conf Series FE			Yes	Yes		
R^2	0.026	0.027	0.085	0.086	0.139	0.139
Clusters	7298	7298	7295	7295	7217	7217
Obs	612103	612103	612100	612100	612022	612022

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered on the conference level. $log(1 + \text{cit})$ for a window of five years is used as the outcome variable. Outcome variables: Forward citations of DBLP conference proceedings / patent families by DBLP items within five years. Mean science citations: 3.45 . We analyze how citations by proceedings received within five years are different for proceedings authored by firms. Additionally, we test whether proceedings authored by a firm-sponsor receive more citations. We include as regressors a dummy indicating whether at least one author is affiliated to a firm, *Firm*, a dummy indicating whether the conference where a conference proceeding is presented is sponsored by a firm, *Sponsor*, and one dummy indicating whether the presenting firm is also a sponsor, *Firm=Sponsor*. In all regressions, we control for year FEs. In columns (3) and (4), conference series FE capture time invariant quality and field differences across conference series. In columns (5) and (6), conference event FE also leave out all variation except within individual events. We find that conference proceedings authored by firms receive on average more citations. The coefficient decreases when controlling for conference series FE and conference event FE, but remains highly significant. This implies that firms tend to present research in conference series of the highest quality, but also within conference series and single conference events, proceedings authored by firms receive more citations. Overall, scientific articles which are associated with at least one firm receive roughly 4.4% more citations than other proceedings in the same conference as seen in column 5. The results on sponsorship suggest that corporate sponsorship is concentrated among high-quality conferences. However, this effect fully stems from firms choosing to sponsor high-quality conference series, rather than individual events within conference series (compare columns 1/2 and 3/4). Since sponsorship is defined at the conference level, it cannot be included in columns 5 and 6. When presenting and sponsoring coincide, the proceeding receives especially many citations. Column 6 shows that within a conference event, firm citations where the firm is also sponsoring receive 11% more citations compared to proceedings where the firm does not also sponsor. These descriptive results do not imply any causality. Possibly, proceedings receive additional attention through the advertising of sponsorship, so that sponsoring creates an additional halo effect which leads to more visibility and follow-on research. This would constitute a causal mechanism and our observations at conferences suggest such a possibility. However, equally likely firms especially sponsor when they expect to also present strong research at a conference.

Table A-4: Heterogeneity: Sponsorship

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
Participation	0.021^{***}	0.112^{***}	0.001	0.097^{***}	0.073^{***}	0.005
	(0.008)	(0.042)	(0.003)	(0.025)	(0.021)	(0.008)
Participation× Sponsor	0.058^{***}	0.190^{***}	0.003	0.108^{**}	0.112^{***}	0.045^{***}
	(0.019)	(0.056)	(0.008)	(0.043)	(0.042)	(0.017)
Sponsor	-0.029^{***}	-0.097^{***}	0.000	-0.049^{**}	-0.051^{**}	-0.022^{**}
	(0.010)	(0.031)	(0.004)	(0.024)	(0.023)	(0.009)
Science citations (L)	0.028^{***}	0.391^{***}	0.003^{***}	0.156^{***}	0.146^{***}	0.036^{***}
	(0.001)	(0.006)	(0.000)	(0.003)	(0.004)	(0.001)
Patent citations (L)	0.008^{***}	0.183^{***}	0.003^{***}	0.194^{***}	0.073^{***}	0.012^{***}
	(0.002)	(0.006)	(0.001)	(0.006)	(0.006)	(0.002)
Research similarity (L)	0.023^{***}	0.234^{***}	0.006^{**}	0.009	0.044^{**}	0.020^{***}
	(0.007)	(0.044)	(0.003)	(0.025)	(0.019)	(0.007)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.076	0.366	0.031	0.185	0.190	0.084
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	16.6	16.6	16.6	16.6	16.6	16.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

Table A-5: Heterogeneity: Participation intensity for citations (linear specification)

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
# Firm's proceedings×						
Participation	0.013***	0.063***	0.003**	0.050***	0.046***	0.007**
	(0.004)	(0.013)	(0.001)	(0.010)	(0.009)	(0.003)
# Firm's proceedings	−0.005**	−0.028***	−0.001	−0.025***	−0.022***	−0.002
	(0.003)	(0.009)	(0.001)	(0.006)	(0.006)	(0.002)
Sponsor×						
Participation	0.080***	0.151***	0.016***	0.181***	0.172***	0.050***
	(0.017)	(0.028)	(0.006)	(0.033)	(0.028)	(0.013)
Sponsor	−0.041***	−0.074***	−0.008***	−0.091***	−0.085***	−0.025***
	(0.010)	(0.015)	(0.003)	(0.018)	(0.016)	(0.007)
Science citations (L)	0.026***	0.383***	0.002***	0.150***	0.138***	0.034***
	(0.001)	(0.005)	(0.000)	(0.004)	(0.004)	(0.002)
Patent citations (L)	0.006***	0.175***	0.002***	0.189***	0.067***	0.011***
	(0.002)	(0.006)	(0.001)	(0.006)	(0.006)	(0.002)
Research similarity (L)	0.014**	0.210***	0.002	−0.002	0.017	0.011*
	(0.007)	(0.031)	(0.002)	(0.020)	(0.018)	(0.006)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.075	0.373	0.026	0.193	0.186	0.083
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	34.6	34.6	34.6	34.6	34.6	34.6

Notes: Linear specification version of A-6. The number of firm proceedings is winsorized at five. * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Standard fixed effects include conference, origin × field, origin × firm, year × origin and year × firm fixed effects.

Table A-6: Heterogeneity: Participation intensity for citations

	(1)	(2)	(3)	(4)	(5)	(6)
	Science cit (present)	Science cit (past)	Patent cit (present)	Patent cit (past)	Collaboration	Hiring
1 × Participation	0.011	0.055	−0.001	0.064***	0.040**	−0.003
	(0.007)	(0.040)	(0.003)	(0.020)	(0.019)	(0.007)
2 × Participation	0.026***	0.164***	0.001	0.101***	0.084***	0.013
	(0.008)	(0.044)	(0.003)	(0.028)	(0.024)	(0.009)
3, 4 × Participation	0.041**	0.226***	0.002	0.168***	0.136***	0.025**
	(0.016)	(0.067)	(0.006)	(0.046)	(0.035)	(0.013)
5+ × Participation	0.061**	0.289***	0.017**	0.237***	0.236***	0.025
	(0.024)	(0.069)	(0.007)	(0.048)	(0.057)	(0.020)
Sponsor, no proceedings × Participation	0.036	0.188*	−0.010	0.090	0.104*	0.021
	(0.028)	(0.102)	(0.010)	(0.062)	(0.053)	(0.030)
Sponsor + Proceedings × Participation	0.099***	0.350***	0.009	0.254***	0.214***	0.065***
	(0.030)	(0.065)	(0.011)	(0.060)	(0.061)	(0.024)
2	−0.005	−0.040***	−0.001	−0.012	−0.015*	−0.007**
	(0.003)	(0.014)	(0.001)	(0.009)	(0.009)	(0.004)
3, 4	−0.011	−0.068**	−0.001	−0.048**	−0.041**	−0.014**
	(0.008)	(0.027)	(0.003)	(0.022)	(0.019)	(0.007)
5+	−0.020	−0.099***	−0.011**	−0.089***	−0.101***	−0.010
	(0.017)	(0.038)	(0.005)	(0.029)	(0.037)	(0.013)
Sponsor, no proceedings	−0.016	−0.073	0.004	−0.016	−0.034	−0.013
	(0.013)	(0.046)	(0.005)	(0.029)	(0.025)	(0.014)
Sponsor + Proceedings	−0.040**	−0.128***	−0.003	−0.079**	−0.067*	−0.032**
	(0.017)	(0.038)	(0.007)	(0.033)	(0.035)	(0.014)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Proceeding-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.083	0.371	0.028	0.196	0.196	0.085
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	5.9	5.9	5.9	5.9	5.9	5.9

Notes: Reports the coefficients underlying table 4 in section 8. * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Standard fixed effects include conference, origin × field, origin × firm, year × origin and year × firm fixed effects.

Table A-7: Conference location × year fixed effects

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
Participation	0.019**	0.095**	0.002	0.096***	0.045**	−0.002
	(0.008)	(0.048)	(0.003)	(0.026)	(0.019)	(0.007)
Science citations (L)	0.029***	0.393***	0.003***	0.158***	0.150***	0.037***
	(0.001)	(0.007)	(0.000)	(0.003)	(0.004)	(0.001)
Patent citations (L)	0.008***	0.183***	0.003***	0.194***	0.075***	0.013***
	(0.002)	(0.006)	(0.001)	(0.006)	(0.005)	(0.002)
Research similarity (L)	0.027***	0.253***	0.006*	0.015	0.075***	0.029***
	(0.008)	(0.050)	(0.003)	(0.026)	(0.019)	(0.007)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard FE	Yes	Yes	Yes	Yes	Yes	Yes
Dest × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.090	0.378	0.033	0.198	0.212	0.088
Observations	5126273	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114	1114
DV cond. mean	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	32.1	32.1	32.1	32.1	32.1	32.1

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.Standard fixed effects include conference, origin × field, origin × firm, year × origin and year × firm fixed effects.

Table A-8: Conference series × origin fixed effects

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
Participation	0.060^{**}	0.202^{*}	0.010	0.135^{**}	0.162^{**}	0.019
	(0.026)	(0.114)	(0.009)	(0.068)	(0.065)	(0.020)
Science citations (L)	0.025^{***}	0.360^{***}	0.002^{***}	0.147^{***}	0.134^{***}	0.034^{***}
	(0.002)	(0.009)	(0.001)	(0.005)	(0.006)	(0.002)
Patent citations (L)	0.006^{***}	0.169^{***}	0.003^{***}	0.186^{***}	0.067^{***}	0.011^{***}
	(0.002)	(0.006)	(0.001)	(0.007)	(0.006)	(0.002)
Research similarity (L)	−0.001	0.179^{**}	0.000	0.002	−0.007	0.011
	(0.018)	(0.086)	(0.006)	(0.050)	(0.047)	(0.015)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Conference FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.072	0.383	0.051	0.207	0.165	0.114
Observations	5112327	5112327	5112327	5112327	5112327	5112327
DV cond. mean	1066	1066	1066	1066	1066	1066
Number clusters	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	9.3	9.3	9.3	9.3	9.3	9.3

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level. Standard fixed effects include conference, origin × field, origin × firm, year × origin and year × firm fixed effects.

Table A-9: Researcher location × Firm × Year fixed effects

	(1) Science cit (present)	(2) Science cit (past)	(3) Patent cit (present)	(4) Patent cit (past)	(5) Collaboration	(6) Hiring
Participation	0.037^{***}	0.149^{***}	0.000	0.139^{***}	0.116^{***}	0.014
	(0.014)	(0.058)	(0.005)	(0.034)	(0.034)	(0.013)
Science citations (L)	0.028^{***}	0.398^{***}	0.004^{***}	0.158^{***}	0.146^{***}	0.036^{***}
	(0.001)	(0.008)	(0.001)	(0.004)	(0.005)	(0.002)
Patent citations (L)	0.007^{***}	0.181^{***}	0.003^{***}	0.196^{***}	0.073^{***}	0.012^{***}
	(0.002)	(0.006)	(0.001)	(0.006)	(0.006)	(0.002)
Research similarity (L)	0.014	0.217^{***}	0.008	−0.020	0.015	0.015
	(0.011)	(0.057)	(0.005)	(0.033)	(0.029)	(0.012)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Firm × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.109	0.415	0.079	0.221	0.223	0.135
Observations	4917944	4917944	4917944	4917944	4917944	4917944
DV cond. mean	997	997	997	997	997	997
Number clusters	0.010	0.158	0.002	0.050	0.048	0.011
F (First)	24.6	24.6	24.6	24.6	24.6	24.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

Table A-10: Full-model specification with various cluster levels

	(1) Science cit (present)	(2) Science cit (present)	(3) Science cit (present)	(4) Science cit (present)
Participation	0.021^{***}	0.021^{*}	0.021^{**}	0.021^{***}
	(0.007)	(0.012)	(0.008)	(0.007)
Science citations (L)	0.029^{***}	0.029^{***}	0.029^{***}	0.029^{***}
	(0.001)	(0.005)	(0.001)	(0.001)
Patent citations (L)	0.008^{***}	0.008^{**}	0.008^{***}	0.008^{***}
	(0.002)	(0.003)	(0.002)	(0.002)
Research similarity (L)	0.025^{***}	0.025^{***}	0.025^{***}	0.025^{***}
	(0.006)	(0.009)	(0.008)	(0.007)
Conf. distance controls	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes
Cluster	Origin	Firm	Origin-Dest	Paper
R^2	0.083	0.083	0.083	0.083
Observations	5126273	5126273	5126273	5126273
Number clusters	1114	3235	88398	235902
DV cond. mean	0.010	0.010	0.010	0.010
F (First)	31.6	144.5	49.6	141.4

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

Table A-11: Science citations (OLS/IV)

	(1) Science cit (present)	(2) Science cit (present)	(3) Science cit (past)	(4) Science cit (past)
Participation	0.005^{***}	0.021^{***}	0.043^{***}	0.113^{***}
	(0.000)	(0.007)	(0.002)	(0.041)
Science citations (L)	0.031^{***}	0.029^{***}	0.402^{***}	0.394^{***}
	(0.001)	(0.001)	(0.004)	(0.006)
Patent citations (L)	0.009^{***}	0.008^{***}	0.188^{***}	0.184^{***}
	(0.002)	(0.002)	(0.006)	(0.006)
Research similarity (L)	0.041^{***}	0.025^{***}	0.307^{***}	0.240^{***}
	(0.003)	(0.006)	(0.012)	(0.043)
Method	OLS	IV	OLS	IV
Conf. distance controls	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes
R^2	0.090	0.083	0.379	0.372
Observations	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114
DV cond. mean	0.010	0.010	0.158	0.158
F (First)		31.6		31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

Table A-12: Patent citations (OLS/IV)

	(1) Patent cit (present)	(2) Patent cit (present)	(3) Patent cit (past)	(4) Patent cit (past)
Participation	0.001***	0.001	0.012***	0.098***
	(0.000)	(0.003)	(0.001)	(0.024)
Science citations (L)	0.003***	0.003***	0.168***	0.158***
	(0.000)	(0.000)	(0.003)	(0.003)
Patent citations (L)	0.003***	0.003***	0.200***	0.195***
	(0.001)	(0.001)	(0.006)	(0.006)
Research similarity (L)	0.007***	0.007**	0.094***	0.012
	(0.001)	(0.003)	(0.004)	(0.024)
Method	OLS	IV	OLS	IV
Conf. distance controls	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes
R^2	0.031	0.031	0.227	0.192
Observations	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114
DV cond. mean	0.002	0.002	0.050	0.050
F (First)		31.6		31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

Table A-13: Collaboration (OLS/IV)

	(1) Collaboration	(2) Collaboration	(3) Hiring	(4) Hiring
Participation	0.015***	0.074***	0.003***	0.006
	(0.001)	(0.020)	(0.000)	(0.008)
Science citations (L)	0.155***	0.147***	0.037***	0.037***
	(0.003)	(0.004)	(0.001)	(0.001)
Patent citations (L)	0.077***	0.074***	0.013***	0.012***
	(0.005)	(0.005)	(0.002)	(0.002)
Research similarity (L)	0.104***	0.047***	0.023***	0.021***
	(0.005)	(0.018)	(0.002)	(0.007)
Method	OLS	IV	OLS	IV
Conf. distance controls	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes
R^2	0.214	0.196	0.087	0.087
Observations	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114
DV cond. mean	0.048	0.048		0.011
F (First)		31.6		31.6

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.

D Interviews: understanding firms' attendance of scientific conferences

In this paper, we capture firms' conference attendance on a large scale through data on authorship of conference proceedings and on sponsorship of conference events. We collected qualitative evidence to better understand the reality behind these indicators. We attended two large, high-quality conferences to gather evidence in support of our assumptions. The first conference was the European Conference on Computer Vision 2018 (ECCV, https://eccv2018.org/) in Munich, Germany. The ECCV is a large biannual $A^{\star}$ conference in computer vision, a subfield of AI. In 2018, about 3500 persons participating. The second conference was the Neural Information Processing Systems conference 2019 (NeurIPS, https://nips.cc/) in Vancouver, Canada, with more than 13000 participants. At ECCV and NeurIPS, we interviewed more than 50 in total, between scientists, HR representatives and engineers, of more than 20 firms and about 20 academic scientists. We talked to firms of several countries and of various size and with different levels of participation. We investigated their activities at conferences and the processes taking place before, during and after conferences. Falling short of a full qualitative study, we here report our general impressions from the interviews. This follows from section 3 of the paper where we describe the main characteristics of CS conference series and the modes of participation of firms.

In summary, firms activities at a conference can be categorized in (i) scientific activities and, (ii) branding and recruiting activities. These two categories relate to rather distinct underlying dynamics and processes. The former is reflected in the conference participation of scientists who present their work and normally interact with their academic and corporate peers. Generally, firm scientists conveyed the impression of a high degree of autonomy, having considerable freedom in the decision of which conferences to attend and, to a large extent, what to present. Firm-level processes, mostly unknown to academic scientists, play a role mostly in the screening of presented work before and in the follow-up activities after the conference. Interestingly, the screening of work submitted to conferences concerns primarily a selection based on quality: most firms have in place internal peer-review systems (and not of hierarchical approval) to ensure presenting above-average scientific work.

Nonetheless, this also entails guaranteeing the presence of sufficient intellectual property protection. All firm scientists interviewed declared that prior to a conference presentation, firm lawyers would verify whether a patent application is necessary, to protect possible valuable inventions and to avoid compromising the future option of obtaining a patent.[21] However, no one declared this to have ever been an impediment to their participation. It also remains that the work of firms scientists presenting at the conference appeared often not directly related to current product development. Some scientists said that the research closely related to product development is normally maintained secret and performed by different organizational units.

[21] In most patent jurisdictions, rendering public an invention generates prior art which jeopardizes the novelty of an eventual patent application also if inventors and authors of the publication are the same

After a conference event, all firms appeared to have in place knowledge sharing processes. Depending on the firm, these may take the form of informal activities, such as the sharing of references among colleagues (also who did not participate to the conference). More often, researchers were expected to write more structured reports or to prepare presentations on the content of the conference to be discussed in internal meetings. In some cases this was supported by an internal IT information system to trace the participation of different individuals to different conferences and events and maintain information on their feedback.

Official recruiting and branding activities are mostly carried out by personnel at the conference booths and are directly connected with sponsorship. HR personnel, in particular, advertise job opportunities, mostly for PhDs and young researchers, attend at the booth to all potential candidates and schedule possible follow-up interviews after the conference. The HR units then take care of preparing the material and define the main activities before the conference, and have follow-up meetings to discuss the outcomes and possible improvements after the conference.

Despite being distinct activities, scientific and branding/hiring activities are not disconnected. On the one hand, the firm sponsorship and the personnel employed at the firm's booths also advertise more generally scientific activities of firms scientists and can offer organizational and logistic support to their scientists. The promotion of research activities performed by firm, especially focused on the specific contributions at the focal conference, is at least equally evident than the promotion of job positions. The sponsorship benefits and "infrastructure" of large sponsors, especially at large conferences like NeurIPS, is used to create opportunities for divulging research results, even to offer specialized tutorials and workshops, beyond the presentation of proceedings or organization of workshops that may be part of the normal conference program.

On the other hand, several HR representatives referred to have experimented also participating to conferences without the presence of scientists. This however proved to be ineffective also for hiring objectives, due to the difficulty of engaging with other scientists. The presence of scientists at the booths is planned, in order to facilitate the conversations with potential candidates for job positions that are often interested in discussing in detail research developed by the firm. And the promotion of research at the firm is clearly complementary to engaging possible candidates. Interestingly, most HR representatives we interviewed declared that the decision of the conferences to sponsor often follows the preferences of where scientists want to present their research. Scientists did not seem to care much about whether the firm sponsored or not an event when deciding where to participate, and would very well, and often do, participate without corresponding sponsorship. Informal connections and interactions of scientists at the conference may also constitute a vehicle to reach and engage candidates. The few sponsors we could talk to with a small booth and no parallel scientific activity demonstrate limited interactions with the conference participants and their booths were poorly attended. One of these sponsors' representative (from a large firm) explicitly expressed dissatisfaction for the lack of a more significant investment by the firm, in her/his own words, "to a community that I deem important for our research units".

The evidence discussed here is necessarily anecdotal. In particular, it is based on only two events and a sample of interviewees necessarily selected by the presence at these conferences. Moreover, the level of investment of firms at ECCV18 and NeurIPS19, similarly to other conference series in ML, has risen sharply in the latest years. Nonetheless, we can very well expect that the type of firm activities carried out at other conferences would be equivalent, and, while the level of investment may have varied over time and across subfields, the nature of these activities would likely be the same. Most importantly, this evidence stands as a proof that the participation of firms to conferences constitute a substantial firm-level investment which is well approximated by our empirical quantitative data.

E Appendix: Data sources

In this appendix we extend the description of our data sources and data construction, following the section 4 of the paper. Table 1 summarizes the type of information obtained from each source and gives references to the source. The relationships between the data sources are visually documented in figure A-5.

Table A-14: Data sources.

Data source	Variables
DBLP	Conference, conference series information including place, time and presented papers, author disambiguation http://dblp.uni-trier.de/
CORE	Conference series quality ranking, sub-fields classification http://www.core.edu.au/conference-portal
WOS, Scopus	Affiliation information, citations, scientific classifications of articles, sponsorship information
SNPL data	NPL citations from patents to conference proceedings Knaus and Palzenberger (2018) and Poege et al. (2019)
PATSTAT	Patent information, applicant and inventor names and addresses
ICAO, BTS	Direct flight connections, Airport regions https://www4.icao.int/newdataplus https://www.bts.gov/
ORBIS, GRID, EU Scoreboards	Firm names, ownership structure, industry information http://www.grid.ac https://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards_en

DBLP has a very broad coverage (Cavacini, 2015) and, compared to other sources, contains more consistent conference and conference series information. Additionally, DBLP supplies a high-quality author name disambiguation (Kim, 2018). DBLP has the highest coverage rate among specialized

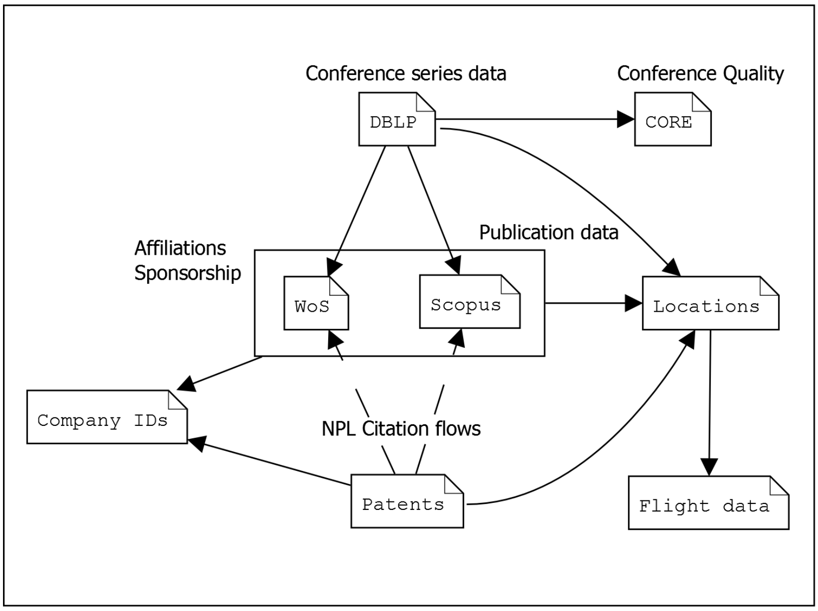


Figure A-5: Structure of the dataset

databases. The data provide an identifier for conference series. Conference event locations and dates are not available as independent fields but can be easily parsed from conference volumes titles. WoS and Scopus have a higher coverage rate, due to the coverage of other fields, but the information in DBLP is more consistent and representative for CS (Cavacini, 2015). Additional information on DBLP is available at dblp.uni-trier.de. A recent discussion of the disambiguation procedures is available at blog.dblp.org/2020/01/08/corrections-in-dblp-2019/.

Other relevant bibliographic information is missing in DBLP, which we obtain from Web of Science (WoS) and Scopus. Since Scopus is available to us from 1996 onwards, we focus our attention to those years. Both WoS and Scopus are widely used bibliometric databases with large coverage of different scientific fields, but possibly with lower coverage of specific fields relative to specialized databases like DBLP. The match between DBLP and the complete WoS and Scopus is done using the DOI and the cleaned title. Matches are verified using page numbers, publication years and author names and only matches showing sufficient overlap are kept. Necessarily, we drop conferences and conference proceedings for which no match is found in WoS or Scopus. We can match up to 90% of the DBLP entries with an item in WoS and/or Scopus.

We add information on conference series quality and CS research subfields from the Computing Research and Education (CORE) data, curated by the Computing Research and Education Association of Australasia. The CORE data classify conference series into the quality-rank levels $A^{\star}$, A, B and C and subfields. The CORE data constitutes an expert-based assessment of conference quality and subfields and is meant to cover a comprehensive set of all relevant conferences in CS. We match CORE to our data manually, partially supported by probabilistic string matching algorithms. We retain exclusively conference series which match with CORE ranking information and drop conference series which are unclassified. We use the latest available version of CORE, which provides the broadest coverage. Consequently, our rank classification is time-invariant. However, by comparing different versions of CORE rankings (2008, 2010, 2013, 2014, 2017, 2018) it is evident that changes in ranks are rare and in most cases minimal.

Table A-15: Observation counts

	Observation counts			
	All	WoS/Scopus	With CORE	≤ 2010
Dataset				
Proceedings	1617817	1444813	982548	612103
Conference Events	22404	20361	10973	7298
Conference Series	3767	3505	1087	1042
Firms				
All Firms		9941	7316	5470
Participants		9173	6791	5042
Sponsors		2121	1398	1027

Notes: Observation counts for different matching steps. Fourth column is the estimation sample. Third column from the right is relevant for the descriptive part. First column: All DBLP items. Second column: DBLP items found in WoS or Scopus. Third column: Also restricting to conference series matched with CORE. Last column: Also restricting to 1996-2010.

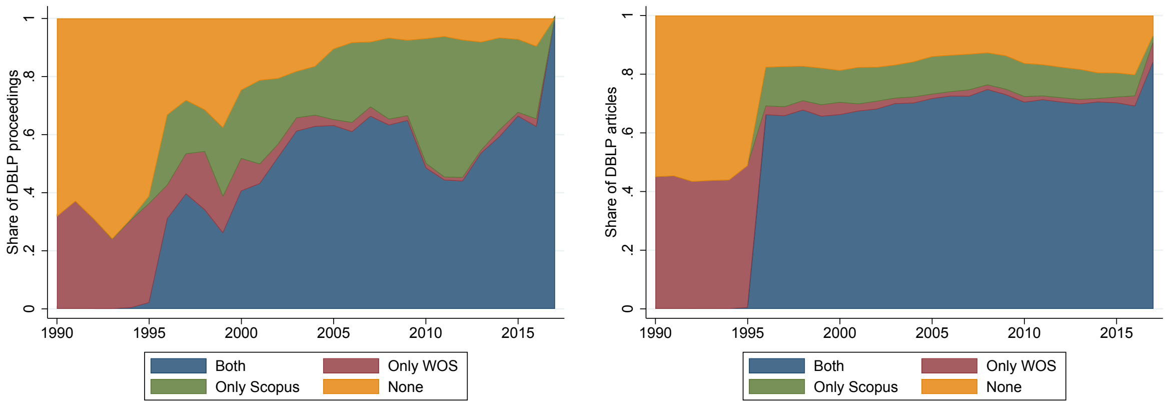
Table A-15 provides an overview of the number of observations in our data. Merging DBLP with WoS/Scopus and CORE inevitably reduces the number of available observations. Thanks to the combination of both, 90% of DBLP (since 1996) is maintained after matching with WoS and Scopus. The achieved coverage rate of DBLP in Scopus and Web of Science is displayed for proceedings in figure A-6a and for articles in figure A-6b, where after 1996, rates of 70-90% are observed. The full Scopus database is only available to us from 1996 onwards, which explains the lack of coverage before. Clearly, without Scopus, the analysis would lack representativeness, but the WOS adds around 10% in all years. This forces us to restrict our period of analysis to after 1996. Combining DBLP, WOS and Scopus guarantees to obtain the largest possible coverage of bibliographic information in CS in this period of time.

The match with CORE data leads to a more substantial drop in the number of unique conference series and conference events originally covered. However, we verified that these are largely small and less relevant conferences, with few corresponding proceedings each. We still retain 982548 conference proceedings, corresponding to 70% of the initial total (the number of proceedings in DBLP matched with WoS or Scopus).

Most importantly, as noted in the paper, our data cover 75% of all conference series listed in CORE. Eighty percent of conference series listed in CORE and not in our data are of the lowest quality rank, C. This implies that the data cover almost the entirety of top and medium ranked conferences in CORE. In general, the sample is biased against small conference events, short-lived conference series, and conference series of the lowest quality, that are less likely covered in a generic bibliographic database as WoS or Scopus, and are less likely ranked in CORE.

We can claim that the data are largely representative of all relevant conference events in CS in our period of observation. Table A-15 also shows the difference between our estimation sample with

Figure A-6: DBLP items covered by WOS / Scopus



(a) Proceedings

(b) Articles

Notes: Shows results of the DBLP-WOS/Scopus match. Match was based on DOI/title, cleaning using page numbers, publication years and author names.

years 1996-2010 and our full sample 1996-2015. Citation-based variables require time windows in which the citations can be observed. We choose five-year windows. This truncation issue forces us to limit to 1996-2010 the sample for econometric analyses. The full dataset (up to 2015) consists of a total of 10973 conference events in the 1996-2015 period – pertaining to 1087 conference series and more than one million proceedings. A total of 7316 firms have participated to at least one conference event, either authoring at least one conference proceeding or sponsoring a conference event. The sample up to 2010 comprises instead 5470 firms, 7298 conference events pertaining to 1042 conference series, and a total of 612103 of conference proceedings. In the remainder of the paper, we also present descriptive statistics limited to this sample.

E.1 Matching firms

We generate a list of firm entities that we use as candidates. Our goal is to provide global coverage of the - probably - most important firms which conduct scientific publishing. For this reason, we use the firm names from the EU scoreboards as well as the firm included in GRID. The Scoreboard lists the by R&D expenditure top companies worldwide. In the first year of the list, 2003, 500 EU and 500 global companies are separately listed. Over time, the length of the lists was increased, so that the 2017 Scoreboard lists the top 1000 EU firms and worldwide the top 2500. The 2017 Scoreboard is the last included in our data. All in all, this adds roughly 8300 distinct firm name strings, of which often several refer to the same firm entity. For GRID, use a snapshot from May 2018. GRID, as a curated dataset of research-active entities is a prime candidate for adding firms likely involved in scientific activities. We only add entities labeled as company, which adds another roughly 21,000 match candidates. We further wanted to complement this list by firms who possibly use information from conferences in their technological activities, but do not publish frequently enough to occur in the curated GRID list. Therefore, we add all firm names for firms which in ORBIS were found to be connected to at least one patent. Also, we added firms from the US and DE section of ORBIS to try to capture smaller firms this way. Especially the latter part expands the set of firms too much by too many irrelevant candidates, so we did not further expand to additional countries.

Matching bibliometric information to firms is particularly hard as little additional information besides the affiliation string exists. Location information is often not given, as is the case for sponsor information. When it is available, it often does not refer to the headquarter location but to the particular research lab. Therefore, we try to enrich the affiliation name with contextual knowledge from the Internet, following the approach by Autor et al. (2020). We search for the affiliation string in a search engine and retain the first ten results. We disregard very frequent occurrences, where for example many firms are listed on a single website. We also use frequency weighting to put a higher weight on less common entries.

The match uses the software package Dedupe. (Gregg and Eder, 2019) Dedupe provides a probabilistic algorithm which, based on manually crafted training data, calculates weights for different input features. These input features are the web search-based similarities, but also traditional string similarity measures. Dedupe also calculates a minimum similarity threshold for which matches are kept. This is done based on a comparison of precision and recall scores. The matching step returns for each affiliation string a set of candidate firm strings which this affiliation string might belong to.

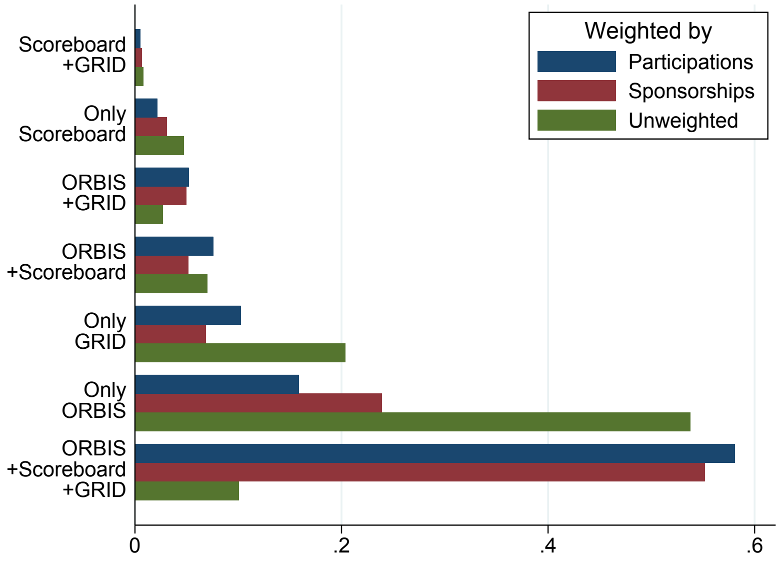
In the next step, we cluster the n:m match provided by Dedupe to group firm strings which belong to the same entity. In GRID, ORBIS and the EU Scoreboards, several possibilities for writing of the same firm name are possible. Additionally, firms may have been renamed, merged, acquired etc. Incidentally, the web search-based algorithm is by itself quite good at picking up these name changes. However, this step required a lot of manual refinement. Whenever multiple entity names were grouped, we validated these choices. When in doubt, the clustering implicitly and our validation explicitly clustered entities in larger groups. So, if two firms merged during a part of the sample time frame, we consider them to be the same entity for our full sample. Also, when the matching algorithm was not able to confidently distinguish subgroups of conglomerates, we grouped them into one entity. This happens with firms like Samsung or LG. The firm clusters yield our firm entities for this study.

Figure A-7 shows the sources of firm observations in the conference dataset. These give an overview over successful matches and of the extent of clustering. The number of firms in each category is weighted by the number of proceedings authored by (blue) and the number of conferences visited (red). Also, an unweighted count is provided. Most individual matches are from ORBIS only, followed by GRID only. However, the most important firms can actually be found in all three databases ("ORBIS+Scoreboard+GRID").

The advantage of our approach is to provide a very global, comprehensive firm dataset. Previous studies have typically only focused on large, listed US companies. Given the degree of internationalization observed in our data, this would substantially underestimate the role of firm science in CS. The disadvantage of not using one consistent firm dataset is that further descriptives at the firm level are difficult to obtain. For example, the GRID data does not contain any further firm-level information, whereas both Scoreboard and ORBIS do. We therefore being with match descriptives based on the Scoreboard only, which provides the most consistent set of firm-level information. Subsequently, we attempt to provide industry classifications for all firms.

We apply our match algorithm for a variety of data sources. At the core of this study is the match between affiliations for conference participants and sponsorship information found in WOS and Scopus. The most relevant of those are for CS conferences also found in DBLP. However, for figure A-1, conferences in other fields are also informative. Further, we use the same matching strategy to match firm applicants from patents citing computer science proceedings or otherwise relevant for computer science (the technology main area 'Electrical Engineering'). Due to this broader match target set, there can be a number of firms which are matched to some affiliation or applicant string but never occur in the computer science dataset.

Figure A-7: Sources of firm data



Notes: Shows the data sources of firms in the dataset. Firm data is taken from the EU Scoreboards, ORBIS as well as GRID. The number of firms in each category is weighted by the number of proceedings authored by (blue) and the number of conferences visited (red). An unweighted count is provided as well (green).

Table A-16 shows, for the subsample of the 2010 Scoreboard, which share of entities can be matched. We focus on one Scoreboard slice as combining several slices would create a distorted sample. Large companies are always retained, whereas small companies would frequently enter and leave the yearly lists. In the 2010 Scoreboard, we can find 69.5% of the Scoreboard companies in any data source, including conferences, journal publications, conferences outside of CS and relevant patents. However, also 69.5% ever participated to a conference. The shares are necessarily higher in some sectors and smaller in others. In Telecommunication Services or Telecommunication Hardware and Equipment, more than 70% of all firms ever participate to a conference. However, in all sectors, some companies show some engagement with the academic community.

Intensity of participation and sponsorship also varies substantially across sectors. Table A-16 shows this in columns 4-7. These columns show the share of Scoreboard firms that ever participated or sponsored a CS conference in the 1996-2015 time period as well as the average number of conferences they participated to or sponsored. As some examples, the sector "Software and Computer Services" contains both smaller IT companies as well as the global players of IT. "Leisure Goods" contains some companies involved in electronic entertainment among a greater number of companies unrelated to CS. This explains the low participation share but high average participation intensity.

Table A-16: Descriptives: Firms at conferences (Scoreboard 2010 subsample)

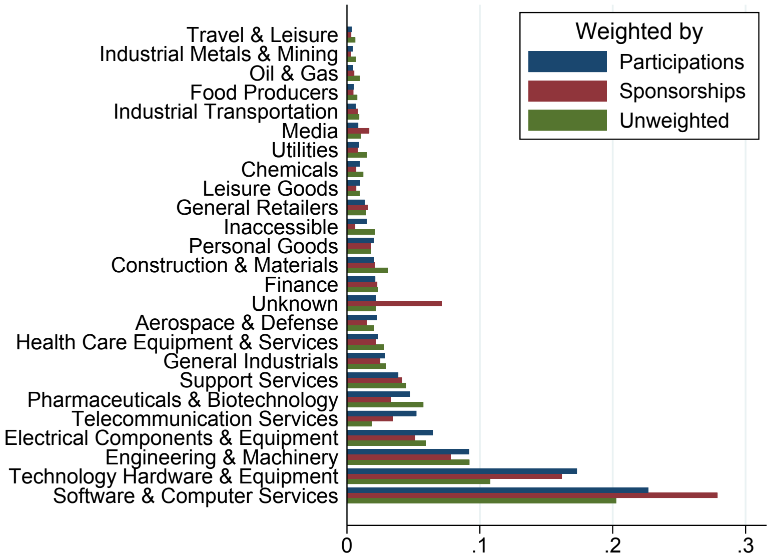
	All	Ever Matched		Participation		Sponsorship	
	N	N	Share	Ever	Total	Ever	Total
Aerospace & Defense	46	37	0.80	0.54	17.52	0.20	0.76
Chemicals	109	91	0.83	0.17	5.14	0.04	0.29
Construction & Materials	69	38	0.55	0.19	1.17	0.04	0.12
Electrical Components	148	108	0.73	0.37	16.01	0.14	1.14
Engineering & Machinery	272	160	0.59	0.31	5.98	0.07	0.27
Finance	80	37	0.46	0.15	0.34	0.06	0.11
Food Producers	67	43	0.64	0.15	0.43	0.03	0.03
General Industrials	56	31	0.55	0.25	13.29	0.11	0.57
General Retailers	20	10	0.50	0.20	1.45	0.00	0.00
Health Care Equipment	73	56	0.77	0.36	1.01	0.04	0.05
Industrial Metals & Mining	36	29	0.81	0.36	1.03	0.00	0.00
Industrial Transportation	15	9	0.60	0.40	0.93	0.00	0.00
Leisure Goods	28	17	0.61	0.32	30.57	0.18	1.82
Media	17	11	0.65	0.41	4.76	0.12	1.18
Oil & Gas	36	33	0.92	0.47	3.44	0.17	0.44
Personal Goods	60	30	0.50	0.13	0.43	0.05	0.10
Pharmaceuticals & Biotech	229	185	0.81	0.13	0.72	0.05	0.06
Software & Computer Services	207	124	0.60	0.43	39.43	0.22	6.00
Support Services	37	21	0.57	0.32	7.14	0.14	0.68
Technology Hardware	253	214	0.85	0.67	30.16	0.23	2.89
Telecommunication Services	26	23	0.88	0.81	109.31	0.54	5.35
Travel & Leisure	26	13	0.50	0.23	1.62	0.08	0.12
Utilities	58	48	0.83	0.33	2.03	0.07	0.10
Total	1968	1368	0.70	0.34	13.5	0.12	1.3

Notes: Shows the number of firms with conference participation and their activities, exemplary for the 2010 Scoreboard. The first column shows the number of firms, by industry. While the Scoreboard contains overall 2000 entries, in some cases multiple parts of conglomerates are listed separately, for example Samsung. In the matching process, these cannot be distinguished with high accuracy and are joined into one entity. The second column shows the number and share of firms that could ever be matched in any target dataset, including CS conferences and journals, conferences outside of CS and relevant patents. The remaining firms could never be matched. Columns four to seven show the share of Scoreboard firms that ever participated or sponsored a CS conference in the 1996-2015 time period. As almost all firms who sponsor also participate, the 'Ever Participation' shares also show shares of firms found in the conference dataset. Columns five and seven show the total number of conferences attended as well as the number of sponsored conferences.

We can also classify firms outside of the Scoreboards into industries. The industries by firms are taken from the respective data sources. In the Scoreboard, the classification according to the 'Industry Classification Benchmark' (ICB) is taken. In ORBIS, the 4-digit NACE2 classification is translated into the ICB classification. For the 2016 Scoreboard, we have a direct correspondence to ORBIS, from which we construct a probabilistic match between NACE2 and ICB. This one is extrapolated for the remaining ORBIS entries. GRID on the other hand lacks any firm-level properties. The website URL is available, however. We use this and website addresses from ORBIS to attempt a linkage between GRID and ORBIS. Conditional on a matching website, we require a close string similarity for a match. With this, we further get ICB information for GRID-only entities. When for one firm cluster, more than one source of industry information is available, all sources are weighted equally. With this, we can describe the industries present in our estimation sample.

Figure A-8 shows a distribution of firms across business sectors. The number of firms in each category is weighted by the number of proceedings authored by (blue) and the number of conferences visited (red). An unweighted count is provided as well (green). As expected, business sectors traditionally associated with computer science such as “Software and Computer Services” or “Technology Hardware” are very important. However, also firms from a variety of other sectors are occasionally present at conferences.

Figure A-8: Industries of firms



Notes: Shows the industries of firms in the dataset. Industry information is taken from the Scoreboards or ORBIS, for GRID via an ad-hoc match with ORBIS. The number of firms in each category is weighted by the number of proceedings authored by (blue) and the number of conferences visited (red). An unweighted count is provided as well (green).

E.2 Constructing scientist-firm biographies

For the collaboration and hiring variables, and for self-citations and scientist counts, we rely on disambiguated scientist profiles and affiliation information combined with firm information. The scientist profiles are taken from DBLP, but for the affiliation information, we rely on data from WOS and Scopus. While for the majority of the data, only a paper-firm link is relevant, here a paper-person-firm link is required. To achieve this, we match at the paper level individual authors from DBLP to individuals authors and corresponding affiliations in Scopus, and if not available, in WoS.

With this, we establish a person-year panel and compute fractional association of individual scientists with firms or academia. Whenever a scientist is associated with a firm on a journal article or conference proceeding, that information is also taken into account. If in a given year a scientist features different affiliations from one or several papers, fractional counts are used. In years where the scientist did not publish, linear interpolations from years before and after are used for the variable on firm size of research investments.

There is a small share of cases where the individual information of affiliation cannot be retrieved. A small part of this issue is due to missing affiliation information. The rest comes from a limitation of WoS, where there is no direct link between the the author list and affiliation list, that they are simply listed uniquely in their order of appearance. For this reason, when available, we prefer information from Scopus, which is essentially complete. We also mitigate this issue as far are possible in WoS: the first affiliation can always be assigned to the first person or papers with only one affiliation can be assigned completely. Still, some cases remain where the information is missing.

F Appendix: Similarity measure

We calculate text similarity scores using the cosine similarity between reduced term frequency–inverse document frequency (tf-idf) values of the cleaned abstracts and titles. In a first step, the abstract is cleaned. Cleaning involves concatenating title and actual abstract, removing copyright statements and replacing special keywords with character strings (2D becomes twod, L2 becomes eltwo, ...). Then, everything which is not a character is replaced with a whitespace. We employ stemming, which reduces flexed forms of words to their stem. We also remove stop words. Of the so-cleaned abstract, we take the 50,000 most frequent tf-idf values of one, two and three-grams. We exclude very frequent terms. We then use a truncated singular value decomposition (SVD) to reduce the dimensionality from 50,000 to 300. This approach is also called latent semantic analysis (LSA). The latter name hints at the purpose - finding dimensions that concisely describe the semantic content of an abstract. Multiple words can have the same meaning and the same word can have several meaning, depending on the context. All in all, this approach generates a procedure which maps an abstract into 300 dimensions. For the tf-idf measure as well as the SVD, it is necessary to take the full body of documents into account in a training stage. For this, we use all 2.6 million DBLP items for which we can find abstracts. Once this training stage is completed, individual abstracts can be analyzed. Finally, the cosine similarity is calculated for two transformed abstracts.

We insert mean and maximum similarity scores as outcome variables into our regression setup from section 6. The theoretical range of the similarity scores is between -1 and 1, but observed values are typically between zero and one. In each firm × year × field group, we observe several similarities when firms have published more than one paper. Within these groups, we take the average and the maximum. When a firm has not published a paper in a given year × field group, we set the similarity score to zero.

Table A-17: Mean similarity scores

Mean Similarity	(1) t-1	(2) t	(3) t+1	(4) t+2	(5) t+3
Participation	-0.008	0.043^{***}	0.018^{**}	0.023^{***}	0.015^{**}
	(0.011)	(0.008)	(0.008)	(0.007)	(0.006)
Science citations (L)		0.011^{***}	0.013^{***}	0.012^{***}	0.013^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
Patent citations (L)		0.007^{***}	0.010^{***}	0.009^{***}	0.010^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes
R^2	0.580	0.422	0.597	0.608	0.622
Observations	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114
DV cond. mean	0.087	0.151	0.090	0.087	0.084
F (First)	28.5	28.8	29.0	29.3	29.3

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.This table shows mean abstract similarity scores of firm papers in $t + x$ relative to the focal paper. Only papers within the same CS field are compared. When a firm did not publish in $t + x$, the mean similarity score is set to zero. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.

Table A-18: Maximum similarity scores

Max Similarity	(1) t-1	(2) t	(3) t+1	(4) t+2	(5) t+3
Participation	−0.010	0.075***	0.029**	0.045***	0.022*
	(0.017)	(0.014)	(0.014)	(0.013)	(0.013)
Science citations (L)		0.026***	0.029***	0.026***	0.027***
		(0.002)	(0.002)	(0.002)	(0.002)
Patent citations (L)		0.019***	0.023***	0.022***	0.023***
		(0.002)	(0.002)	(0.001)	(0.001)
Conf. distance controls	Yes	Yes	Yes	Yes	Yes
Conf Ser FE	Yes	Yes	Yes	Yes	Yes
Origin × Field FE	Yes	Yes	Yes	Yes	Yes
Origin × Firm FE	Yes	Yes	Yes	Yes	Yes
Year × Origin FE	Yes	Yes	Yes	Yes	Yes
Year × Firm FE	Yes	Yes	Yes	Yes	Yes
R^2	0.685	0.541	0.697	0.706	0.716
Observations	5126273	5126273	5126273	5126273	5126273
Number clusters	1114	1114	1114	1114	1114
DV cond. mean	0.165	0.245	0.172	0.168	0.162
F (First)	28.5	28.8	29.0	29.3	29.3

Notes: * $p < .1$, ** $p < .05$, *** $p < .01$ Standard errors in parenthesis, clustered at the researcher region level.This table shows maximum abstract similarity scores of firm papers in $t + x$ relative to the focal paper. Only papers within the same CS field are compared. When a firm did not publish in $t + x$, the mean similarity score is set to zero. Firm-proceeding level dataset, where some proceedings were actually at the conference (Participation=1) and some were at another conference (Participation=0). Participation is instrumented by the direct flight availability between the researcher location and the conference location. Firm-proceeding controls include whether the firm cited previous work by the authors in the years before the conference (Science/Patent citations L) and the average abstract similarity between proceedings published by the firm in the previous year and the focal proceeding (Research Similarity L). Dependent variable mean is for actually presented proceedings. Conf. distance controls include the distance between researcher and conference location (log), whether that distance is zero and whether the two locations are in the same US state or non-US country.