NBER WORKING PAPER SERIES

THE DIFFUSION OF NEW TECHNOLOGIES

Aakash Kalyani
Nicholas Bloom
Marcela Carvalho
Tarek Alexander Hassan
Josh Lerner
Ahmed Tahoun

Working Paper 28999
http://www.nber.org/papers/w28999

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2021, revised August 2024

We thank audiences at the Applied Machine Learning webinar, Atlanta Fed, Auburn University, Babson College, the Bank for International Settlements, Baruch College, Bocconi University, CKGSB, Columbia University, College de France, Dartmouth University, Duke University, Durham University, ETH, the Federal Reserve Board, Georgia State University, Harvard University, the Korea-American Economic Association, the London Business School, the London School of Economics, Michigan State University, Northwestern University, Nova Business School, New York University, the Ohio State University, the Royal Bank of Australia, Stanford University, the Toulouse Network on Information Technology, the University of British Columbia, the University of California at San Diego, Santa Barbara, and Santa Clara, the University of Chicago, the University of Maryland, the University of Michigan, the University of Minnesota, the University of North Carolina, the University of Southern California, the University of Texas, the University of Washington, Yeshiva University, and the 2021 NBER Summer Institute, the Fall 2021 NBER EFG meeting, and the 2022 Society for Economic Dynamics and the 2024 American Economics Association annual meetings for helpful comments. Special thanks go to Lisa Kahn for sharing data, Bledi Taska for help on BGT data queries, Gaétan de Rassenfosse, Shane Greenstein, Ben Jones, and Chad Syverson for excellent discussions, and Peter Donets, William Hartog, and Jared Simpson for excellent research assistance. We thank Scarlett Chen, Nick Short, Corinne Stephenson, and Michael Webb for assistance in conceptualizing and researching early versions of this project. Funding for this research was provided by Harvard Business School, the Institute for New Economic Thinking, the Kauffman Foundation, the Sloan Foundation, the Toulouse Network on Information Technology, and the Wheeler Institute. Bloom and Lerner have received compensation from advising institutional investors in venture capital funds, venture capital groups, and governments on venture capital topics. All errors and omissions are our own. The views expressed herein are solely those of the authors and do not necessarily reflect those of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2021 by Aakash Kalyani, Nicholas Bloom, Marcela Carvalho, Tarek Alexander Hassan, Josh Lerner, and Ahmed Tahoun. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Diffusion of New Technologies
Aakash Kalyani, Nicholas Bloom, Marcela Carvalho, Tarek Alexander Hassan, Josh Lerner, and Ahmed Tahoun
NBER Working Paper No. 28999
July 2021, revised August 2024
JEL No. O31,O32

ABSTRACT

We identify phrases associated with novel technologies using textual analysis of patents, job postings, and earnings calls, enabling us to identify four stylized facts on the diffusion of jobs relating to new technologies. First, the development of economically impactful new technologies is geographically highly concentrated, more so even than overall patenting: 56% of the most economically impactful technologies come from just two U.S. locations, Silicon Valley and the Northeast Corridor. Second, as the technologies mature and the number of related jobs grows, hiring spreads geographically. But this process is very slow, taking around 50 years to disperse fully. Third, while initial hiring in new technologies is highly skill biased, over time the mean skill level in new positions declines, drawing in an increasing number of lower-skilled workers. Finally, the geographic spread of hiring is slowest for higher-skilled positions, with the locations where new technologies were pioneered remaining the focus for the technology's high-skill jobs for decades.

Aakash Kalyani
Federal Reserve Bank of St Louis
1 Federal Reserve Plaza
St Louis, MO 63102
aakashk@bu.edu

Nicholas Bloom
Stanford University
Department of Economics
579 Jane Stanford Way
Stanford, CA 94305-6072
and NBER
nbloom@stanford.edu

Marcela Carvalho
Harvard Business School
Soldiers Field Road
Boston, MA 02163
mmello@hbs.edu

Tarek Alexander Hassan
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and NBER
thassan@bu.edu

Josh Lerner
Harvard Business School
Rock Center 214
Soldiers Field Road
Boston, MA 02163
and NBER
jlerner@hbs.edu

Ahmed Tahoun
London Business School
26 Sussex plc, Regent's Park
London NW1 4SA
United Kingdom

The dataset constructed as part of this paper, as well as relevant code, is available at www.techdiffusion.net

1. Introduction

Economists have long recognized that the development of novel technologies is inexorably linked to economic growth. Many studies have sought to understand whether the benefits from adopting new technologies accrue primarily to inventors, early investors, highly skilled users, or to society more widely through, for instance, employment and income growth.[2] Substantial concerns remain, however, as to the implications of new technologies, including whether they contribute to income inequality (e.g., do technology-enabled jobs spread beyond college graduates?) and regional inequality (do technology jobs spread outside Silicon Valley?).[3]

One key obstacle to resolving these questions is that it has proven difficult to measure the development and spread of multiple technological advances in a single framework and to systematically identify those innovations that affect jobs and businesses. In this paper, we use the full text of millions of patents and job postings and hundreds of thousands of earnings conference calls to make progress on this issue. We develop a flexible methodology that allows us to determine which (sets of) technological innovations most affected businesses over the past two decades, trace these back to the locations and firms where they emerged, and track their diffusion through regions, occupations, and industries over time. We then use our newly created data to establish key stylized facts about the development and diffusion of new technologies across space and skill levels.

The first step of our analysis is to develop a methodology for systematically identifying one, two, and three-word phrases (unigrams, bigrams, and trigrams) associated with new technologies through a series of systematic rules, whose robustness we verify through various diagnostic tests. To this end, we intersect information from multiple large corpora of text. First, we use the full text of U.S. patents with application years between 1976 and 2014 to isolate phrases that appear in multiple patents but did not exist before 1970. That is, we isolate new language specific to influential innovations made in the past 40 years. Second, we search for these phrases in Wikipedia to identify which of these new phrases are primarily associated with pages describing new technologies, as opposed to newly recognized problems (such as "climate change") or new management terms (such as "performance metrics"). This procedure identifies 1,899 new

[2] See, for example, Katz and Murphy (1992), Krusell et al. (2000), Piketty and Saez (2003), Autor et al. (2008), Goldin and Katz (2009), Acemoglu and Autor (2011), and Song et al. (2019).
[3] See Tyson and Spence (2017) and Vance (2022) for popular articulations of such concerns.

technology phrases, which we can group into 1,286 unique Wikipedia pages describing new technologies. We refer to these groups of new technology phrases as "technologies."

After establishing our list of new technologies, we then identify patents and job postings that mention these technologies. We use patent inventor addresses to identify the locations where each of the technologies was developed and patent application years to pinpoint the year in which the technology experienced the first large acceleration in patent references (its "emergence year"). We then cross-reference our list of technology phrases with the full text of online job postings to identify 51 million jobs advertised between 2010 and 2019 that mention these new technologies. These granular data uniquely allow us to track the spread of new technologies along a dimension of crucial importance to policymakers: jobs. In particular, we examine the evolution of the number, location, and skill requirements of job postings associated with these new technologies.

In a final step, we use the full text of earnings conference calls held by listed firms between 2002 and 2019 to flag those new technologies that are frequently referenced in these important conversations between firm executives and investors. The most frequently mentioned technologies include "cloud computing," "smart phone," and "machine learning."

Strikingly, the right tail of technologies with the most earnings call mentions also account for the lion's share of the variation in our job-postings and patenting data. For example, the 276 technologies with more than 100 mentions in earnings calls (the top 22 percent) also appear in 39 million job postings (or about 77 percent of all job postings mentioning any new technology), and 33.1% of patents granted by the U.S. Patent and Trademark Office (USPTO) with application years between 1976 and 2014. In this sense, the innovations that feature prominently in managers' discussions also have the largest impact on patents and job postings. We therefore pay special attention to these most economically impactful technologies throughout our analysis.

Our key results are as follows.

First, the locations where new technologies are developed are geographically highly concentrated This concentration is particularly pronounced for the most economically impactful technologies: 33.3% of patents mentioning any new technology and 42.1% of patents mentioning a new technology with more than 100 mentions in earnings calls emerge from just five urban areas: San Jose, San Francisco, New York, Seattle, and Boston.

Based on early patenting activity around the time of each technology's emergence year, we identify which urban areas housed the majority of early patenting for each of our new technologies. We term these urban areas "pioneer locations."

Again, these pioneer locations for new technologies are highly concentrated, particularly so for the most economically impactful new technologies. Collectively, 56.3% of these most impactful technologies come from just two U.S. locations, Silicon Valley and the Northeast Corridor.[4] (Locations in California collectively host a remarkable 41.0% of pioneer locations of these most impactful new technologies.) This extreme concentration is particularly important because new technologies alter the composition of the local job postings in their pioneer locations for several decades, as we show below.

Second, despite this highly skewed initial distribution of pioneer locations, as technologies mature and the number of new jobs related to them grows, they gradually spread geographically. Our favored measure of geographic concentration, the coefficient of variation of the share of jobs associated with a new technology across the 917 core-based statistical areas (CBSAs) in the U.S., falls by 18.5% in the first decade after its emergence. Nevertheless, the implied years to full dispersion across CBSAs is more than 50 years, well beyond the horizon of most policymakers.

Third, while initial hiring is heavily biased towards high-skilled jobs, the mean required skill level of the jobs associated with new technologies declines over time, reflecting a broadening of the types of jobs that adopt a given technology. Specifically, we estimate that, in the year of the average technology's emergence, 57.1% of the initial jobs relating to this new technology require a college degree – a substantial skill bias relative to 30.3% of respondents in the 2015 ACS[5] that hold one. This gap declines by 0.23 percentage points per year, so that 30 years after a technology's emergence year, on average about 50.2% of job postings relating to it still require a college degree.

Fourth, low-skill jobs associated with a given technology spread out across space significantly faster than high-skill jobs, which tend to remain concentrated for long periods of time within the pioneer locations that originally developed the technology.

[4] We define the San Jose-Sunnyvale-Santa Clara and San Francisco-Oakland-Hayward CBSAs as Silicon Valley and the Northeast Corridor as New York-Newark-Jersey City, Boston-Cambridge-Newton, Washington-Arlington-Alexandria, and Philadelphia-Camden-Wilmington.

[5] This share is rising over time, so for example in the 2021 ACS this proportion is 36.4%.

A key implication of these patterns is that new technologies appear to yield long-lasting benefits for the pioneer locations where they were originally developed. These locations host a disproportionate share of high-skilled jobs relating to these new technologies for about four decades after their year of emergence. In short, this concentration of innovation in a handful of urban centers engenders large and persistent regional disparities in economic opportunity, giving a handful of U.S. locations a lasting advantage in high-skill job postings.

To shed light on the mechanisms underlying these patterns, we study the context in which the new technology is mentioned within a given job posting. We find that much of the regional spread of new technologies is driven by low-skill jobs associated with their *use*, whereas jobs related to their *research, development, and production* (RDP) remain persistently concentrated in and around their pioneer locations. That is, pioneer locations that initially developed a technology retain a long-term advantage, because they retain the technology's RDP for long periods of time.

We also show evidence to suggest that some of the observed skill broadening of new technologies is driven by standardization that allows for the use of these technologies by lower-skill workers as the technology matures. By contrast, training and experience with the new technology do not appear to be major drivers of this process.

We conduct a large number of robustness checks, replicating our main results using a wide range of different variations. For example, we repeat our analysis with phrases of different lengths (such as unigrams and trigrams), conduct a human audit of technology phrases and technologies, and use alternative methods for pinpointing pioneer locations and emergence years. Throughout all of these variations, our main findings remain unchanged.

We note three main caveats to our interpretation. First, all of our results regarding jobs rely on the analysis of job postings. In this sense, they measure the characteristics of open positions, but not necessarily the characteristics of the jobs that get filled. Second, by its very nature, our data speak to job openings relating to novel technologies but not to the possible destruction of existing positions by these technologies. Finally, a third concern is what Merton (1968) termed "obliteration by incorporation": when a technology becomes so widely diffused that it is no longer mentioned specifically in job postings. For this reason, we focus on relatively recent technologies, rather than ones, such as electricity or air-conditioning, that have been around for so long that they became normalized.

Our work builds on a large literature that studies the relationship between technology and labor markets. One strand of this literature studies the diffusion of technology. This literature has focused on patterns in a single specific (though important) new technology, from computers (Autor et al., 2003) to broadband (Akerman et al., 2015) to robots (Acemoglu and Restrepo, 2020) to artificial intelligence (Agrawal et al., 2019; Webb, 2020). Other studies have focused on specific innovations during important historical episodes (Griliches, 1957; Goldin and Katz, 1998; Squicciarini and Voigtlander, 2015; Caprettini and Voth, 2020).[6] Comin and Hobijn (2004, 2010) characterize the diffusion of 15 technologies across 166 countries, employing a variety of measures of technological utilization at the country level.[7] We contribute to this literature by identifying hundreds of new technologies, pinpointing their geographic origins, and tracking their spreads across job postings, skill levels, and geographies within the United States.

A second strand is the literature on technology and inequality. Many of these works have sought to estimate the skill bias of technical progress (e.g., Katz and Murphy, 1992; Krueger, 1993; Berman et al., 1994; Autor et al., 1998; Goldin and Katz, 2008; Autor et al., 2008; Michaels et al., 2014; Song et al., 2019). The near-universal approach in this literature is to infer an increase in the demand for skilled labor over time from changes in observed wage differentials – in effect, documenting a change in the economy's aggregate production function. Our work complements this literature by observing this skill bias of technical progress directly: for instance, 57.1% of early jobs involved with new technologies require a college degree.[8] [9]

Closely related, Caselli (1999), Acemoglu et al. (2012), and Acemoglu and Restrepo (2018) study theoretically the forces that drive automation, the substitution of capital for labor, and inequality. A key result in this literature is that balance in the "race between man and machine" arises endogenously if the use of new technologies spreads from high-skill to low-skill occupations over time. We contribute by providing the first direct evidence that this skill broadening indeed occurs systematically for a broad range of technologies. In addition, this theoretical literature argues that

[6] Recent work has examined the importance of supply and demand factors for the speed of diffusion (e.g., Popp, 2002; Acemoglu and Linn, 2004; Greenstone et al., 2010; Moser et al., 2014; Moscona, 2020; Arora et al., 2021). Mokyr (1992) and Gordon (2016) trace out the impact on economic development of a range of great inventions.

[7] A large, related literature studies the role of trade and multinational production in facilitating the diffusion of technology. Recent examples include Buera and Oberfield (2020) and Lind and Ramondo (2022).

[8] Notably, Goldin and Katz (1998) show that the introduction of new manufacturing processes during the early 19th century increased the demand for skilled labor. Krueger (1993) shows that workers who use computers at work earn higher wages.

[9] Van Reenen (1996) and Kline et al. (2019) study how rents from innovation are shared with employees. We relate to these papers by showing evidence that the economic opportunities stemming from the development of new technologies distribute highly unevenly across space, as opposed to across different actors within a given firm.

one mechanism underpinning skill broadening is that new technologies evolve over time into a standardized form that can more readily be used by less educated workers. We provide empirical evidence supporting this standardization mechanism.

A third broad literature examines clustering in entrepreneurial activity and innovation. A number of papers have highlighted persistent advantages in entrepreneurship (Glaeser et al., 2015) and innovation (Moretti, 2021) that certain urban areas enjoy and highlighted mechanisms such as employee mobility across new ventures (Gompers et al., 2005) and localized knowledge spillovers (e.g., Jaffe et al., 1993). We contribute to this literature by providing a systematic approach to identifying and studying pioneer locations. We characterize their distribution across the United States and show there is a general relationship between successful innovation, early employment in a new technology, and the long-term advantage that these locations enjoy in high-skill employment.

Finally, our work adds to a growing literature in economics using text as data. A number of recent papers have used newspaper articles, patents, and firm-level communications to measure concepts that are otherwise hard to quantify (e.g., Hoberg and Phillips 2016; Baker et al., 2016; Hassan et al., 2019, 2021; Bybee et al., 2020; Handley and Li, 2020; Flynn and Sastry, 2020; Kelly et al., 2021; and Sautner et al., 2023). We focus primarily on the full text of job postings, which has received relatively less attention.[10] Important papers by Kogan et al. (2022) and Autor et al. (2024) intersect information from patents, Census job titles, and task descriptions to measure complementarities between innovations and jobs. Our work adds to this literature by introducing a flexible methodology for analyzing the origin and spread of innovations by intersecting multiple large corpuses of texts.

The remainder of this paper is structured as follows. In Section 2, we discuss how we identify and characterize new technologies in the data. Section 3 studies the spatial concentration of the development of new technologies. In Section 4, we explore the diffusion of activity across regions and the associated mechanisms. We present our analysis of the skill-broadening results in Section 5. Section 6 examines diffusion across occupations, industries, and firms. Section 7 presents robustness checks. The final section concludes the paper.

[10] A notable exception is the work by Abis and Veldkamp (2020), who use job descriptions to identify financial analysis positions that leverage machine learning.

2. Identifying and Characterizing Technological Innovations

Our first objective is to identify a list of phrases describing *influential technological innovations developed since 1976*.

We use the term *technological innovation* in the sense of Schmookler (1966) and Jewkes et al. (1969), who distinguish technological from scientific innovation – the former being a set of specific and applied techniques, products, and processes (our focus here) while the latter is a set of general principles. This motivates our use of patents (as opposed to scientific research papers) as a text source.[11] We further distinguish technological from managerial knowledge. While Syverson (2011) and Bloom et al. (2016) argue that managerial rather than technological knowledge can account for substantial differences in total factor productivity across firms, we deliberately focus *on technological but not managerial knowledge* when we require that the new language we isolate from patents describe technologies.[12] By *influential* and *developed since 1976* we mean those innovations mentioned repeatedly in highly cited patents and those that went through a major acceleration in patenting activity after 1976.

We now describe in more detail how we operationalize these concepts in the data.

a. Step 1: Identify phrases associated with influential innovations

We begin by examining patent filings with the USPTO. By law, patents must describe their technological innovation and (at least some) key ways in which it is applied.[13] Because of the importance of the U.S. market, inventors worldwide typically file important discoveries with the USPTO.[14]

We collect all utility patents awarded to U.S. inventors with application years between 1976 and 2014, a total of approximately three million patents. We focus not just on the front page of the

[11] The U.S. patentability standard requires an invention not to be obvious "to a person having ordinary skill in the art" (35 U.S.C. 103), an abstract idea, a law of nature, nor a natural phenomenon (35 U.S.C. 101). See the discussions, for example, by the Supreme Court in *Alice Corp. v. CLS Bank International*, 573 U.S. 208 (2014) at 216 and *Mayo Collaborative Servs. v. Prometheus Labs.*, Inc., 566 U.S. 66 (2012) at 71.

[12] The OECD's Oslo Manual (2005) elaborates on this distinction, providing many examples of what would and would not be included in the two categories.

[13] This requirement is stipulated in the legal concept of "reduction to practice," 35 U.S.C. 112(a).

[14] About half of all patent applications to the USPTO are filed by residents of foreign countries (USPTO, 2020). This pattern reflects the fact that patent protection in any nation depends critically on having a patent issued in that specific nation. Important discoveries (the focus of our analysis) are therefore disproportionately likely to be filed in major patent offices worldwide (Lanjouw et al., 1998).

award, which has been the focus of much of the earlier analytic literature, but on the entire text of these patents. Representative parts of a patent are reproduced in Appendix Figure 1. For more details on this collection process refer to Section 1.1 of the Data Appendix.

To reduce the dimensionality of this voluminous body of text, we remove stop words (such as "of," "the," and "from") following Kelly et al. (2021) and Gentzkow et al. (2019) and represent each patent's remaining text by a vector of all two-word combinations ("bigrams") that appear at least twice in the patent, leaving us with 17 million unique bigrams. In our main specification, we focus on bigrams because they are less ambiguous than single-word keywords. For example, while words like "autopilot" or "cloud" could have a variety of colloquial meanings, "autonomous vehicle" and "cloud computing" are much less ambiguous (e.g., Tan at al., 2002; Bekkerman and Allan, 2004). In Section 7 (robustness), we show that our results extend readily to including unigrams (one-word) and trigrams (three-word combinations), though unigrams generally appear to produce noisier results and trigrams add little to the analysis once bigrams are accounted for.

We next seek to isolate those bigrams that are novel and associated with influential innovations. First, we focus our attention on bigrams associated exclusively with *novel* innovations by dropping "non-novel" bigrams that were in common use before 1970. To this end, we select all text dating prior to 1970 from the *Corpus of Historical American English*, a representative sample of text constructed by linguists from prominent sources (Davies, 2009) that reflects everyday use of English up to 1970. We pre-treat this text in the same way as the patent text, eliminating stop words and extracting bigrams. We then remove any bigram appearing in the *Corpus* (for instance, "equipment used") from our list of bigrams obtained from patents, leaving us with 1.5 million exclusively "novel" bigrams.[15]

[15] At the same time, if the individual words appear in the Corpus, but not in conjunction with each other (e.g., "artificial" and "intelligence" separately, but not as a bigram), we do not delete the phrase.

Second, to identify bigrams associated with *influential* innovations, we retain only those novel bigrams that appear in patents accumulating a total of at least 1,000 patent class and year-normalized citations.[16, 17] This leaves us with 36,563 novel and influential bigrams from patents.

b. Step 2: Identifying technological innovations using Wikipedia

A review of these novel and influential bigrams from patents suggests they fall into three broad categories. Some describe technological innovations, such as "fingerprint sensor," "monoclonal antibody," or "OLED display." Others refer to new (or increasingly visible) problems, such as "greenhouse gases" or "Parkinson's disease." Yet others refer to areas that may have seen substantial new developments or management attention but are not new technologies, such as "account management" and "performance metrics." (Appendix Table 1 shows examples.) As discussed above, we want to focus on bigrams in the first category, not the other two.

To isolate bigrams describing *technological* innovations, we employ Wikipedia entries. We first match each novel and influential bigram to a Wikipedia page by entering it into the Wikipedia search engine and selecting the highest-ranked entry if it mentions the bigram either in the title or the summary or it mentions the bigram at least 10 times in the body of the entry. Bigrams that do not meet these criteria (those without a Wikipedia page) are deleted.

The second step exploits the standardized nature of Wikipedia page entries. Entries describing technological innovations tend to feature sections containing the words *application(s)*, *use(s)*, *type(s)*, *operation*, *characteristic(s)*, *feature(s)*, *device(s)*, *technical*, and *commercial* in their titles. (Appendix Figure 2 provides examples of two Wikipedia pages with these features.) By contrast, pages dedicated to new problems or management innovations tend to feature sections and/or titles that contain the words *responses*, *mitigation*, *problems*, *causes*, *signs*, *symptoms*, *adverse effects*, *management*, *manager*, *risk assessment*, *business model*, *distribution model*, *customer*, *strategy*, and *service provider*. To focus on bigrams associated with technological innovations, we thus

[16] Following Lerner and Seru (2022), normalized citations for a patent p are calculated as: $\frac{Citations_p}{Avg_{\tau,t}(Citations_{p'})}$. $Citations_p$ is the number of citations is received as of 2018 by a patent filed in four-digit Combined Patent Classification (CPC) technology class τ in year t. $Avg_{T,t}(Citations_{p'})$ is the average number of citations received by all patents filed in technology class τ in year t.

[17] For computational reasons, it is necessary to limit the analysis to a subset of the 1.5 million novel bigrams before cross-referencing with other corpuses (steps 2-4). However, where exactly we draw the boundary between influential and non-influential bigrams (1000 normalized citations) has little effect on our results, as discussed in Section 7.

retain only those that are matched with a Wikipedia page with at least one section from the former list, but none of the latter.

This algorithm returns a list of 4,277 bigrams associated with influential technological innovations, which we can conveniently group by the 2,746 unique primary Wikipedia pages that they are associated with. For ease of reference, we refer to these bigrams as "technology bigrams" and their groupings as "technologies," which we label by the Wikipedia page's title.[18] Appendix Table 1 provides examples of bigrams that passed and failed this Wikipedia filtering. For further details on scraping and processing Wikipedia pages, refer to Section 1.3 of the Data Appendix.

c. Step 3: Characterizing technologies using patents and earnings calls

To learn more about when and where each technology was developed, we next cross-reference our list of technology bigrams with our corpus of patents.[19] First, to obtain a measure for each technology's age, we calculate for each bigram the first episode of accelerated patenting. In particular, we first calculate the number of cite-weighted patents (normalized as described in Section 2.a) mentioning the bigram filed in each calendar year. Due to the variability of the patent counts, we smooth the series by taking a centered five-year moving average. Finally, we mark the first year in which (a) the technology reaches 100 cite-weighted patents and (b) the next five years had at least 10% annual growth in the (smoothed) weighted patent filings. For ease of reference, we refer to this year as the bigram's "emergence year."

This process is illustrated in Figure 1, which depicts the time series and the emergence year for four technology bigrams. Digital video, for instance, emerges in 1986, as the time series grows by at least 10% for five consecutive years through 1991. Using this definition, we assign an emergence year after 1976 to 1,899 technology bigrams (1,286 technologies). The remaining bigrams exhibit no single five-year period of accelerated growth in our sample, and thus predominantly describe older technologies (such as diesel fuel and whey protein). In Section 7, we

[18] To check the accuracy of this procedure, we conducted a formal human audit following the methodology in Baker et al. (2016). To this end, we developed a detailed coding guide to train three research assistants on the definition of new technologies given above. We asked them to each manually classify a random sample of Wikipedia pages matched to one of our 35,563 novel and influential bigrams from patents. Collectively, the research assistants coded 700 entries. The research assistant's coding of bigrams that were confidently technological (with a confidence greater or equal to three out of five) corresponded to the answer of the Wikipedia filter 73% of the time. In addition to this human audit, we run robustness tests using a manually reviewed sample of technologies in Section 7.

[19] When cross-referencing our technology bigrams with patents and other corpuses, we generally allow for all forms of the bigram, including singular, plural, and concatenations. We require the bigram to appear at least twice in the patent.

show our results are robust to using a range of other plausible approaches to defining each bigram's emergence year. The key is simply to obtain some meaningful distinction between older and newer innovations.

Second, to identify regions pioneering the early development of a technology, we identify the CBSAs that collectively account for a majority of early patents mentioning the technology. In particular, for each technology bigram, we calculate the number of patents in each CBSA within the first ten years of the bigram's emergence year. We then sort the CBSAs by the number of patents mentioning that bigram and denote those CBSAs with the most of these patents that collectively account for at least 50% of the total patents mentioning the technology bigram in this period as "pioneer locations." Thus, if the top three CBSAs accounted for 35%, 25%, and 8% of the patents containing a bigram in this period, the first two would be coded as pioneer locations.

Third, we can gauge the extent to which a given technology poses economic challenges or opportunities to incumbent firms by cross-referencing our list of technologies with the full text of 321,373 corporate earnings calls held by 11,905 listed companies and compiled by Refinitiv EIKON between 2002 and 2019. Publicly traded firms hold quarterly earnings calls to discuss results and the companies' prospects. These calls (and the transcripts that we analyze) consist of a presentation by management (typically the chief executive and/or chief financial officer) and then questions posed by investors and analysts with answers by the executives. They have been shown to be indicators of some of the most important issues facing these organizations (Bushee et al., 2003; Matsumoto et al., 2011; Hassan et al., 2019, 2021).[20] To gauge the extent to which each technology features in the conversations at these listed firms, we record the number of unique earnings calls in which each of our technologies is mentioned.

Table 1 gives a flavor of these data. It shows the top technology, as measured by the number of earnings calls mentioning it, by year of emergence of the technology, as well as its associated bigrams. Top technologies emerging in the late 1970s and early 1980s include the hard disk drive, barcode reader, and personal computer. The mobile phone emerges in 1985, followed by digital video and debit cards. The 1990s brought machine learning and the hybrid electric vehicle. The top technologies from the 2000s include the smartphone, social networking, and the self-driving car. Taken together, these technologies appear to accurately reflect the changing nature of

[20] Some examples of mentions of bigrams in earnings calls are shown in Appendix Table 2.

technological innovation over the past decades. Appendix Table 3 lists all new technologies that are mentioned in more than 100 earnings calls. While we make no claim of completeness, we argue they constitute perhaps the most representative sample of economically impactful technological innovations constructed to date.

Table 2 provides examples of the pioneer locations for several technology bigrams. For example, pioneer locations for machine learning (a technology that emerged in 1994 according to our measure) are New York, Seattle, San Jose, and San Francisco, whereas digital imaging's pioneers are Rochester (Kodak's headquarters), San Jose, San Francisco, and Fort Collins (the longtime home of Hewlett Packard's desktop and peripherals business).[21]

These steps illustrate how, once we have identified a list of new technologies and their associated phrases, we can build a rich panel dataset of these technologies by identifying their mentions in other text sources. We expand on this theme next.

d. Step 4: Cross-reference with job postings

We finally cross-reference our list of technologies with the full text of online job postings, which we source from Burning Glass (BG). BG aggregates online job postings from online job boards (such as indeed.com), employer websites, and other sources into a de-duplicated database.

We employ two datasets from Burning Glass. The first is a standardized dataset (used recently by Hershbein and Kahn, 2018; Deming and Noray, 2020; and Atalay et al., 2020), where each de-duplicated job posting is geo-coded and assigned to a Standard Occupational Classification (SOC) code and a North American Industry Classification (NAICS) code.[22] The second dataset has thus far received less attention by researchers. It contains the raw unprocessed text of the job postings, which we use to identify jobs involved with the research, development, production, or use of our technologies. Appendix Figure 3 displays some representative pages from a full BG database entry.

[21] Appendix Table 4 shows, for selected states, the technology where the state most dominated early innovation; that is, the technology where the state contributed the largest share of early patenting. The table also shows intuitive patterns. For example, Massachusetts accounts for 13.6% of the early patenting in the technology "antibody-drug conjugate," and similarly, Michigan accounts for 49.9% in "electronic stability control."

[22] We make extensive use of the former, which are available for 80% of all postings. Industry classifications are available for a more limited 41% of postings. We use industry data only in Section 6. The strings with firm names are available for 66% of all postings.

We have data from BG for all available years, 2007 and 2010-2019, a total of roughly 200 million job postings. We drop 2007 jobs from our baseline analysis because Burning Glass is missing data for 2008-09, though including the 2007 data has little impact on our results.

Our analysis of job postings thus focuses on the diffusion of technologies with emergence years post-1976 in job postings in the 2010s. That is, technologies with an emergence year of 1980 are thirty years old by the time we see them diffusing in job postings, whereas technologies with an emergence year of 2005 are five years old, and so on. For this reason, we are careful to highlight any differences in the variation across technologies vs. within technologies over time in our analysis below.

We associate each posting with a skill level, location, industry, and firm as follows (for details, see Section 1.5 of the Data Appendix): *Skill level.* We construct a skill level for each six-digit SOC code (the most detailed level) given in BG by measuring the share of persons with a college degree, the share of persons with a PhD or a master's degree, the average wage, and the average years of schooling in the American Communities Survey (ACS 2015 release), using the respondents who report their occupation in that six-digit SOC code.[23] *Location.* We use the county names provided by BG to uniquely assign job postings to one of the 917 CBSAs in the United States. *Industry:* We allocate a job posting to an industry using the four-digit NAICS code provided by BG.[24] *Firm:* To allocate job postings to firms, we extend the methodology of Autor et al. (2020) and cluster employer strings associated with job postings together on the basis of top search results on Bing.com. For more details on the firm mapping, please refer to Section 3 of the Data Appendix.

To identify job postings associated with each technology bigram, we simply check whether the job posting mentions that bigram and create an indicator variable that is equal to one if it does:

$$Technology\ Job_{i,\tau,t} = 1\{b_\tau \epsilon\ D_{i,\tau}\}, \quad (1)$$

where b_τ is a given technology bigram τ associated with one of our new technologies and $D_{i,\tau}$ is the set of bigrams contained in job announcement *i* posted in year *t*. In our main specification, we exclude the first and last 50 words of the job posting from this set to avoid picking up mentions of

[23] For SOC codes in job postings where we do not find any persons surveyed in the ACS, we match them to the closest available SOC code in the ACS. For example, data for SOC Code 38-1967 were not available, so we match these observations to 38-1960. In total, the dataset includes 837 SOC codes.
[24] NAICS codes typically have six nested levels; the four-digit level is referred to as "industry group."

the technology in the initial firm description or ending boilerplate language, as opposed to the task to be performed by the employee, as we discuss below.

To interpret what it means for a job posting to mention a technology, we conduct a human audit of 1,000 randomly selected technology job postings (see Appendix Table 7 for details). As expected, the vast majority of mentions relate to a task to be performed by the employee (91% if we trim the first and last 50 words, 80% otherwise). That is, job postings usually mention technologies when the job involves using, producing, or otherwise interacting with the technology. For example, a job ad with mention of "touch screen" (see Appendix Table 7) requires the worker to use a touch screen to enter data. The remaining mentions are either unspecific (4% in our human audit), for example, mentioning that these technologies are available in the workspace, or refer to the company but not the job (4% if we trim the first and last 50 words, 16% otherwise).

For each of our 1,286 technologies, we thus have its year of emergence, a list of pioneer locations where the technology was invented, and a highly granular dataset of job announcements (indexed with a location, industry, occupation, skill level, firm, and year) that involve using, producing, or otherwise interacting with the technology. Most of our analysis focuses on aggregations of these granular data to the technology-time and the technology-location-time levels. Appendix Table 8 provides summary statistics for each level of aggregation. However, the data also open the door for much more granular analyses of job postings for specific firms, locations, and occupations, as we discuss below (see Appendix Table 6 for an example).[25]

Of course, each of the four steps of our data construction can be implemented in different ways, which we highlight when exploring robustness in Section 7. For example, we may choose different thresholds for a technology's emergence year, include or exclude unigrams or trigrams, and employ various human audits of the technologies identified by our algorithms. While each of these variations result in a slightly different sets of technologies and bigrams, we find they have little effect on our main findings below. It should be noted that a number of studies have used employment data from other sources that we do not explore here to understand the diffusion of

[25] Comfortingly, the share of job postings within a given occupation that mentions a new technology with an emergence year post-1979 correlates closely with the share of new job titles created within that occupation since 1980, as identified by Autor et al. (2023) and shown in Appendix Figure 11.

technology. Among the most important of these are Tambe and Hitt (2012), Tambe (2014), and Tambe et al. (2020), who measure the skills of U.S. IT workers using resumes from Linked In.[26]

e. Technologies and earnings calls

Figure 2 shows a binned scatterplot of the number of mentions in earnings calls over the number of job postings mentioning each of our 1,286 technologies. It shows two important patterns: First, both variables are highly correlated – the same new technologies that occupy the discussions of managers and investors in earnings calls are also most frequently mentioned in job postings. The R^2 of a fitted regression line is 57.0%. Second, both distributions are heavy tailed (note the logarithmic scale on both axes), so that a relatively small number of technologies drives the vast majority of the mentions in both job postings and earnings calls. The 276 technologies that are mentioned in more than 100 earnings calls ($EC \geq 100$) account for about 39 million job postings (or about 77 percent of all job postings mentioning any new technology). On average, each of these technologies is mentioned in 141,634 job postings and 7,682 patents.[27]

For ease of reference, we sometimes refer to this highly prolific group of new technologies as "economically impactful" new technologies, in the sense that these new technologies take a significant amount of airtime in earnings calls, and feature prominently in both job postings and patents. It includes all the examples from Table 1 (e.g., smartphone, machine learning, hybrid vehicles).

The figure also shows examples of other, less influential, technologies. Those with between 10 and 99 mentions in earnings calls include the pulse oximeter and the liquid chromatograph. On average, each such technology is associated with 26,128 job postings and 4,287 patents. The group

[26] Tambe (2014) shows that firms that based in regions with considerable number of workers trained in big data skills experience faster productivity growth, an effect that diminishes as these technologies mature. Below, we show the generalizability of this dissipation result and its slow pace.

[27] Interestingly, there is also a clear positive relationship between the numbers of industries in which a given new technology is mentioned and overall earnings call mentions, suggesting that more impactful technologies also tend to be more "general purpose," in the sense that they are relevant for multiple industries.

with under ten earnings call mentions includes the ultrasonic horn, suction filtration, and NMOS transistors, with on average 1,165 job postings and 3,005 patents.[28,29]

3. Spatial Concentration of New Technologies

We first describe the spatial distribution of innovative activity associated with our new technologies. Table 3 examines the regional concentration of patents that mention new technologies. It shows two major stylized facts.

First, relative to the distribution of the population and the educated workforce, the development of new technologies is regionally concentrated. Of the 917 CBSAs, the top five collectively account for 33.3% of patents mentioning a new technology. As such, the development of new technologies is significantly more concentrated than the distribution of college graduates (22.5%) and the overall workforce (18.9%), but also similar to the concentration of overall patenting activity in the United States (32.4%).[30]

Second, this concentration increases significantly as we condition on increasingly economically impactful technologies as proxied by mentions in earning calls. Panel A in Table 3 shows that the share of the top-5 CBSAs in patents mentioning new technologies with more than 100 mentions in earnings calls is 42.1%. These prolific CBSAs are San Jose, San Francisco, New York, Seattle, and Boston.

Figure 3 shows the share accounted for by these five prolific CBSAs increases monotonically from 24.6% of patenting relating to relatively low-impact technologies (those mentioned in zero or one earnings calls) to 46.5% of the highest-impact group (new technologies mentioned in 500+

[28] Note that our notion of technology as an "applied technique, product, or process" naturally recognizes the NMOS transistor as a separate technology from the smartphone, even though the latter might contain or even require the former. Similarly, in the context of job postings, there is a clear distinction between a job task requiring use of a smartphone and a job task involving NMOS transistors. In this sense, we are using language, which naturally generates different terms for different technologies that workers and firms interact with, to measure a technology's economic importance in job postings and earnings calls. These notions of economic importance are thus also quite distinct from broader notions of scientific importance, where understanding electricity and transistors are prerequisites to building smartphones.

[29] Appendix Figure 4 shows the average number of job postings for each category of technology.

[30] These totals are each for the five CBSAs highest on that individual measure. Only one of the largest CBSAs for patents – New York – is on the top five list for employment, highlighting how population size is not the primary correlate of patenting share.

earnings calls). In short, *the most commercially impactful innovations also have the most geographically concentrated origins.*[31]

Interestingly, to preview our results below, this extreme concentration of economically impactful innovation is the only significant difference that we document between more and less economically impactful innovations. Aside from their concentrated origins, less impactful technologies appear to evolve and spread similarly to their more impactful counterparts.

In the same vein, Figure 4 shows the distribution of pioneer CBSAs – the urban areas that account for a majority of *early* patenting of economically impactful technologies (again, those with more than 100 earnings call mentions, $EC \geq 100$). Panel A of Figure 4 presents these patterns in map form; and Panel B presents them in a bar chart showing CBSAs' share of all pioneer locations. In Panel B, we combine San Jose and San Francisco as Silicon Valley, which accounts for 28.7% of all pioneer locations. Jointly, all California CBSAs account for about 41.0% of all pioneer locations. Major cities in the Northeast Corridor, New York, Boston, Washington DC, and Philadelphia, jointly account for 27.6% of all pioneer locations. The top two clusters alone – Silicon Valley and the Northeast Corridor – thus account for 56.3% of all pioneer locations. This result highlights the high concentration of the most economically impactful innovative activity within America over the last decades.

These pioneer locations tend to have highly educated workforces and a high density of university activity. For each CBSA- technology pair (e.g., "smart phone" and the San Jose CBSA), Appendix Figure 5 presents binned scatter plots of patents mentioning each technology in the ten years prior to the emergence date (per capita, normalized by total CBSA population) and regional characteristics. In all cases, there is a strong association between measures of education/university presence and per capita patents relating to new technologies. Interestingly, these associations are significantly more pronounced when we condition on economically impactful new technologies. Regions with a greater research university presence or a more educated workforce are thus significantly more likely to be involved in the early development of key new technologies.[32]

[31] Appendix Table 5 also reports the coefficients for analyses using the top five, three, and one CBSA(s), as well as similar analysis partitioning technologies by the number of associated job postings.

[32] This finding matches the large literature on the geographical concentration of innovation and its connections to university activity, such as Jaffe (1989), Jaffe et al. (1993), Zucker et al. (1998), and Furman and MacGarvie (2007). Moretti (2021) illustrates these effects by examining inventor moves to larger innovation clusters, showing that they experience significant increases in inventive productivity. (This result was hinted at in Forman et al. (2016) as well.)

We show evidence below that this concentration of innovation in a handful of urban centers engenders large and persistent regional disparities in economic opportunity, as measured by job postings in local labor markets. In this sense, a handful of U.S. locations appear to have a comparative advantage in developing technologies that most impact firms and labor markets.

4. Diffusion across Regions – Region Broadening and Pioneer Advantage

We next seek to understand the diffusion of new technologies in job postings across regions.

To understand the geographic spread of technology job postings, we define the normalized share of job postings in CBSA c mentioning a technology bigram τ in year t:

$$Normalized\ share_{c,\tau,t} = \frac{\sum_{i \epsilon c} Technology\ Job_{i,\tau,t} \,/\, \sum_{i} Technology\ Job_{i,\tau,t}}{\#Jobs_{c,t}/\#Jobs_{t}}. \qquad (2)$$

The numerator measures the share of all jobs relating to a given technology τ at a given point in time *t* that are located in c; and the denominator is the share of location c in the overall U.S. labor market at *t*. $Normalized\ share_{c,\tau,t}$, therefore, measures the regional over- or under-representation of job postings associated with each technology bigram relative to the distribution of overall open jobs. Values above one denote over-representation and below one under-representation.[33]

In Figure 5, we present a series of maps displaying the spread of job postings mentioning economically impactful new technologies ($EC \geq 100$). The blue circles identify the same pioneer locations as in Figure 4, but now superimpose purple dots that show the intensity of the normalized share of job postings relating to these new technologies 0-5, 6-10, 11-20, and 21-30 years after the technology's year of emergence. Darker dots correspond to a higher normalized share of jobs.

Two patterns stand out. First, as time goes by, jobs relating to new technologies gradually spread across space (region broadening). Second, there is a remarkable alignment between the CBSAs that pioneer early development in technologies and the CBSAs that host their early employment. Even after accounting for differences in the size of the local labor market, early employment is strongly concentrated in the same places where the technology was originally developed (pioneer advantage). We next substantiate these two patterns formally.

a. Region broadening

[33] Throughout, we cap this variable at the 99th percentile of non-zero observations.

We first examine the overall geographic dispersion of technology job postings. To this end, we calculate the coefficient of variation of the normalized share of technology job postings by dividing the standard deviation of $Normalized\ share_{c,\tau,t}$ across locations c in year t by its mean in year t for each technology bigram τ.[34] If technologies are uniformly spread out across CBSAs, then the normalized share takes a value of 1 for each CBSA, and the coefficient of variation calculated across CBSAs is 0.

The average coefficient of variation in our sample of new technologies is 4.69, which suggests that technology job postings relating to these new technologies are on average highly concentrated compared to, for example, the coefficient of variation for the normalized share of the local population that holds a college degree (2.90).

Using a regression framework, Table 4 examines the evolution of this coefficient of variation over the technology's life cycle. Panel A of this table reports results from regressions of the form:

$$CV_{\tau,t} = \alpha_0 + \beta_{RB}(t - t_{0,\tau}) + \delta_\tau + \varepsilon_{\tau,t}, \quad (3)$$

where $CV_{\tau,t}$ is the coefficient of variation across CBSAs for technology bigram τ in year *t*, and $(t - t_{0,\tau})$ is the number years since emergence of technology bigram τ in year $t_{0,\tau}$ (capped at 30 years, given we have little data for technologies older than 30 years). δ_τ denotes a full set of technology bigram fixed effects, which we constrain to sum to zero, so that the intercept α_0 measures the average coefficient of variation in the year of emergence.[35] The slope coefficient, β_{RB}, measures the speed of decay of this concentration with each passing year since emergence. Panel A, column 1 reports estimates for economically impactful technologies ($EC \geq 100$). Columns 2 and 3 report results for all new technologies, without and with bigram fixed effects, respectively. Throughout, we cluster standard errors at the technology level (the unit of observation is a technology bigram). [36]

[34] Appendix Table 8 summarizes the data used in this and subsequent regression analyses.

[35] Because the coefficient of variation, as well as some of the other constructed moments used in the following tables, become noisy with insufficient data, we take steps in the regressions to down-weight technologies that are mentioned in relatively few job postings. First, we weight observations by the square root of the total number of job postings mentioning that technology, capped at 100, meaning that technologies with more than 10,000 postings receive full weight, while those with less than 10,000 postings are weighted by their square root. Second, we exclude technology bigrams with less than 1,000 job postings. In practice, these adjustments have little impact on our estimates (see Section 7).

[36] Due to the linear form of our estimating equation, the within-technology-and-time variation is effectively degenerate, so that we cannot simultaneously introduce technology and time fixed effects. In this sense, there is no way of distinguishing cohort from time effects, as is common in such analyses (see Hall et al., 2007). However, note that the dependent variable is already normalized to

In column 1, we find that on average, in the year of emergence, the coefficient of variation is 5.58 (significantly greater than 0). With each additional year since emergence, this coefficient of variation decreases by 0.068 (s.e.=0.026) points (or 1.22%). Taking these estimates at face value suggests that technology job postings on average take 82 years to fully disperse across the U.S. (This latter projection is of course considerably out of sample.)

Figure 6 shows this pattern graphically using a binned scatterplot: a technology's job postings are geographically highly concentrated in the early years after its emergence. Within 30 years, this geographic concentration drops by about a third (36.6%). Interestingly, the figure also shows this process of spread, measured in the pooled set of technologies, is close to linear in the data.[37]

In column 2 of Panel A of Table 4, we show this pattern is almost identical when we include all new technologies in our sample. Appendix Table 19 tests explicitly for differences in the rate of spread between technologies with fewer and greater than 100 earnings call mentions, finding no economically significant differences.

In column 3 of Panel A of Table 4, when conditioning only on within-technology variation, we find a somewhat faster rate of spread. With each additional year, the coefficient of variation falls by 0.153 (s.e.=0.012) or 1.85% – implying 54 years to full dispersion.

Panel B of Table 4 shows similar results (following the same specification as column 3 of Panel A) using alternative measures of geographic concentration as dependent variables: the mean normalized share of a technology's job postings in the top five CBSAs relative to the mean across all CBSAs, the percentage of CBSAs with a normalized share of a technology's job postings of less than 10% (that is, the representation of CBSAs with almost no activity associated with that bigram), and the sum of squared deviations of the normalized share from one (similar to the Herfindahl-Hirschman Index). The consistent pattern is for a slow decline of concentration, however measured: all three measures fall with time, but again imply time periods in excess of 50

account for any time trends in the overall coverage of job postings: By construction, the coefficient of variation of the overall job postings in our database is 0 for all *t*, meaning that our measure of the diffusion of technology job postings is immune to variations over time in the share of jobs covered by BG or the shares of regional labor markets covered.

[37] Appendix Table 9 includes a quadratic term and shows it is indistinguishable from zero. Additional specification tests suggest the relationship is closer to linear than log in the data. Consistent with the literature on S-curves, which studies the speed of adoption of a given technology (Griliches, 1957; Rogers, 1962), we do find significant concavity in the rate of spread when conditioning only on variation *within* technologies (column 3 of Appendix Table 9).

years to full dispersion. This is a strikingly slow rate of convergence, given that the typical political cycle is around five years, and most Americans work for less than 50 years.

b. Pioneer advantage

Table 5 formally explores the second pattern: pioneer advantage. We quantify the advantage that pioneering regions (CBSAs that account for a majority of the initial patenting in a technology) retain in that technology's job postings, even as region broadening occurs. Panel A reports results from the specification:

$$Normalized\ share_{c,\tau,t} = \alpha_0 + \beta_P\ Pioneer_{c,\tau} + \beta_D\ Pioneer_{c,\tau}(t - t_{0,\tau}) + \delta_c + \delta_\tau + \delta_t + \varepsilon_{c,\tau,t} \quad (4)$$

where $Pioneer_{c,\tau}$ is a dummy variable denoting the pioneer status of the CBSA; $\delta_c, \delta_\tau, \delta_t$ denote CBSA, technology bigram, and year fixed effects respectively. Columns 1 and 2 examine job postings relating to economically impactful technologies ($EC \geq 100$); columns 3 and 4 show results for all technologies.

In column 1, we see that pioneer locations enjoy a significant pioneer advantage on average: The normalized share of technology job postings is 31.1 percentage points higher in its pioneer locations on average throughout the sample period. Column 2 shows that this advantage is much larger in the year of emergence (108.4 percentage points), but then decreases significantly over time -- on average by 3.2 percentage points per year or 3.0% (0.032/1.084). The initial advantage of the pioneering locations for job postings relating to the economically impactful technologies they develop thus lasts for decades, with an implied 34 years to zero advantage.[38]

In column 3, we include all technologies and again find an almost identical pattern – albeit with a somewhat larger point estimate for the pioneer advantage in the year of emergence of 1.321 (s.e.=0.254). In column 4, we look at technology job postings in the neighborhood of pioneer locations by adding a dummy for CBSAs within 100 miles of a pioneer location, $Pioneer\ Neighbor_{c,\tau}$, and its interaction with the number of years since the emergence of the technology. The estimates suggest that some of the pioneer advantage spills over to these adjacent communities, with a 15.8 percentage point higher normalized share in the year of emergence.

[38] In Appendix Table 10, we test the robustness of our results to the addition of interacted fixed effects. We find that decay rates of pioneer advantage are similar across these specifications.

Again, this advantage appears to decay over time, though the decay is not statistically distinguishable from zero.

c. Mechanisms

Given the extreme regional concentration of new technologies' pioneer locations, and the long-term advantage in jobs these regions appear to enjoy, a key question is *why* this advantage appears to be so persistent. We take two steps to better understand the mechanisms behind this persistence: First, we examine the skill requirements of the jobs spreading across space. Second, we analyze the words around those in which the new technology is mentioned in the job posting to learn about whether the job is involved with developing or using the new technology.

Pioneer advantage in high- vs low-skill jobs

We first analyze differential rates of spread of high- versus low-skill jobs relating to new technologies. To compute a job posting's skill requirement, we use the 6-digit SOC code allocated to the job posting by Burning Glass and assign it the average level of college education respondents report in the 2015 ACS for that occupation.[39, 40]

Columns 1 and 2 of Panel A of Table 6 report results from the specification:

$$\log\left(CV_{\tau,t}^{s}\right) = \alpha_0^s + \gamma_1^s\left(t - t_{0,\tau}\right) + \delta_\tau + \varepsilon_{\tau,t},\ s \in \{H, L\} \quad (5)$$

where $CV_{\tau,t}^{s}$ is the coefficient of variation of the normalized share of technology job postings across CBSAs, as in Section 4.a, calculated separately for $s\epsilon\{H, L\}$ – high-skill jobs (H) and low-skill jobs (L). For the purposes of this exercise, we define high-skilled jobs as those which are classified in occupations with more than a 60% college-educated share in the 2015 ACS (28.4% of all jobs on BG) and low-skilled jobs as those with under a 30% share (42.5% of all jobs). Column 1 reports results for high-skill jobs, and column 2 reports results for low-skill jobs. To facilitate the direct comparison of differential rates of spread between these two types of jobs, we take logs of the dependent variable so that the slope coefficient is now directly informative about the percentage decline in the coefficient of variation per year.[41]

[39] As an example, Appendix Table 11 shows the list of top occupations by share of job postings for some of our top technologies (see Section 2 of the Data Appendix for details).
[40] The BG data also includes an indicator for a college requirement for a subset of observations. However, since this subset is quite limited, we prefer using SOC codes to generate this variable.
[41] Results are almost identical when using a tripartite division of skill levels, as Appendix Table 12 shows.

We find that the geographic concentration of low-skill jobs (in column 2) decreases 1.1 percentage points or 41% (0.038/0.027 - 1) faster than that of high-skill jobs (in column 1). This difference is statistically significant at the 1% level, as we report in the label of Table 6 (and in the labels of subsequent analyses where we can compare coefficients across equations). Figure 7, Panel A shows this differential decay graphically, this time also including across-technology variation (without technology bigram fixed effects). Again, low-skill technology job postings spread at a significantly faster rate.

This pattern is similarly prominent when analyzing pioneer location advantage. In columns 1 and 2 of Panel B of Table 6, we repeat the regression specification in column 2 of Table 5, but now separate between high- and low-skill jobs (all definitions are as above). We find that the pioneer advantage in a technology's job postings is significantly more persistent for high-skill jobs than for low-skill jobs. While the former decays at 2.2 percentage points per year, the latter erodes at a faster 3.2 percentage points. These estimates imply it takes 45 years for a pioneer location's advantage in high-skill jobs to erode, whereas that for low-skill jobs lasts only 31 years.[42]

Taken together, this evidence suggests that the overall geographic spread of technology jobs is driven by low-skill jobs, while high-skill jobs take significantly longer to spread across space. That is, the pioneer locations involved with the early development of a technology tend to retain a significant and very long-lasting advantage in high-skill job postings relating to that technology.

Research, Development, and Production

While there are a number of hypotheses that can be offered for these patterns, our text-based methodology allows us to look carefully at one leading explanation: the movement of new jobs from technology research, development, and production (RDP) to technology use.

The text in the job announcements contains rich information to distinguish these two types of jobs. For example, a job posting involved with a technology's RDP might state "*you will be designing the graphics module for our* ***virtual reality*** *training system,*" while one involved with a technology's use might read "*the role will involve assisting customers and selling tickets from your* ***smart tablet*** *in the entrance of the cinema*." (Additional examples in Appendix Figure 6.)

[42] Both results are again almost identical when we repeat these estimations for the subset of technologies with EC≥100 in Appendix Table 20.

To systematically identify the cases that involve RDP of new technologies, we use an iterative procedure that combines an unsupervised learning algorithm with some human judgment to identify word patterns associated with RDP job postings. The first step is developing a set of plausible keywords (generated by the authors) that are commonly used when describing positions relating to the RDP of new technologies ("research," "and develop," "and development," "customization of," "to build," and "to design"). We then use an embedding vector algorithm trained on earnings calls to identify other phrases (unigrams and bigrams) that are typically used in similar context to these keywords – in effect, using the embedding model like a custom-trained thesaurus.[43] For each of these suggested phrases, we examine ten excerpts from job postings to check for false positives. We then add to our initial list those suggested phrases that had at least eight true positives (no more than two false positives). After updating the list, we go through the steps again iteratively – now asking the embedding model for phrases proximate to the union of already selected phrases – until we have exhausted all useful suggestions that meet the threshold of eight out of 10 true positives. Appendix Table 13 lists the full set of selected phrases.

Using this classification, we systematically flag all job postings that mention a new technology within 15 words of one of our RDP keywords and categorize all others under "use." To verify the accuracy of the resulting classification, we conduct a human audit of 1,000 randomly sampled technology job postings. We assign team members to read and classify these job postings into either RDP or use of the associated technology. In this random sample, we are able to correctly classify 63.1% of technology RDP postings and 68.1% of technology use postings. With this distinction in hand, we calculate the coefficient of variation of the normalized share of technology job postings for each technology and year separately for the RDP and use job postings.

Columns 3 and 4 of Table 6 (Panel A) examine the differential spread of these two different types of technology job postings, estimating region broadening separately for each group. Again, we see large differences: technology-using job postings spread out 157.1% (=0.036/0.014-1) faster than postings that involve technology RDP jobs. This difference is again significant at the 1% level.

[43] Specifically, we use the Word2Vec Python package Gensim trained on earnings calls (sourced as noted above) from 2002 to 2019. For the training process, we used the default parameters: 200 dimensions, ignoring words that appear fewer than 50 times, and a context window of 15 words. We train on earnings calls, instead of job postings, because this type of language model tends to perform poorly when trained on short texts.

We find a similar pattern for the pioneer advantage in RDP job postings. Columns 3 and 4 of Panel B in Table 5 re-estimate regression specification (4) and calculate the advantage of pioneer CBSAs in technology job postings that involve the RDP and use of new technologies. We find that pioneer advantage in job postings involving the use of new technologies is smaller initially (with a constant term suggesting 158.1% more such jobs in the pioneer location in the year of emergence) and dissipates significantly over time (-0.030, s.e.=0.012). By contrast, RDP job postings are more concentrated in pioneer locations initially (197.4% higher in the year of emergence), and the decay rate is statistically indistinguishable from zero (though negative and in a similar range as other estimates in the table). (See also Figure 7, Panel B.)

Taken together, these findings suggest that technologies remain highly concentrated in their research, development, and production in the original pioneer location, using highly skilled employees for these activities, but spread out in their application, where lower-skilled employees are utilized. To consider the example of smart phones, these continue to be developed primarily in Silicon Valley by Masters- and PhD-level employees, but jobs involving their use have spread out across the U.S., including positions for sales, repair, maintenance, and utilization, often undertaken by non-college-educated employees. That is, pioneer locations that initially developed a technology appear to retain a long-term advantage in high-skilled jobs, because activities relating to the technology's RDP remain in that location for long periods of time.

5. Skill Broadening

We next turn to examining the skill bias of technology job postings over time. We find a significant high-skill bias in new technologies initially. Over time, the share of lower-skilled job postings mentioning the technology increases, albeit at a relatively slow rate.

We compute the average skill requirement of job postings associated with a particular technology bigram at a point in time by examining the occupational composition of these job postings:

$$Skill_{\tau,t} = \frac{\sum_o N_{o,t}^{\tau} \chi_{o,2015}}{\sum_o N_{o,t}^{\tau}} \quad (6)$$

where $N_{o,t}^{\tau}$ is the number of Burning Glass job postings mentioning bigram τ that are in SOC code *o* at time *t*, and $\chi_{o,2015}$ is the average skill level for occupation *o*, as measured by the 2015 ACS.

For example, if for a technology bigram τ in year t all associated job postings are in an occupation o, then its skill level is equal to the average skill level of workers in occupation o in the ACS.

Table 7 uses a regression framework to describe the evolution of the skill level of job postings associated with new technologies. The specification is identical to equation (3):

$$Skill_{\tau,t} = \alpha_{0,SB} + \beta_{SB}(t - t_{0,\tau}) + \delta_{\tau} + \varepsilon_{\tau,t}\,, \tag{7}$$

but now we use the average skill required for jobs associated with technology bigram τ in year t as the dependent variable. The intercept $\alpha_{0,SB}$ denotes the average skill level of the technology's job postings in its year of emergence, $t_{0,\tau}$. The slope (β_{SB}) denotes this skill level's average speed of decay with each passing year since emergence. Column 1 of Panel A reports results for economically impactful new technologies. Columns 2-4 again include all new technologies.

In column 1, we find that, on average, 57.1% of job postings mentioning a new technology require a college degree in the year of emergence of the technology. As such, jobs associated with a new technology are significantly skill biased, particularly when compared with the share of the U.S. workforce that holds a college degree – about one third. At the same time, this skill content of a technology's job postings is significantly downward sloping over time. With each additional year since emergence, it falls by 0.228 (s.e.=0.092) percentage points on average, implying a rate of skill broadening of 0.40% (=-0.228/57.078) per year.

Figure 8 shows this evolution graphically using a binned scatterplot. Although the pattern of skill broadening is clearly visible, it is worth noting that 30 years after the year of emergence, the average college requirement is still 50.2%, far above the average rate of college attainment in the U.S. population, as noted above. In this sense, new technologies persistently generate a disproportionate share of employment opportunities for high-skill workers for very long periods of time. Column 2 shows almost identical results for the broader sample with all technologies.

One possible concern with these results is that the types of jobs advertised online (as opposed to in printed newspapers) could be changing over time.[44] To address this concern, column 3 shows

[44] Appendix Figure 9 describes the overall volume and the composition of Burning Glass (BG) job postings over time. Panel A shows that BG job postings have increased about one-to-one with job postings captured in the U.S. Bureau of Labor Statistics' Job Openings and Labor Turnover Survey (JOLTS). Panel B shows that the average skill level associated with BG job postings has fallen over time at about 0.7% per year. Panels C and D show that the volume of BG job postings by occupation (pooled across years and by year) is associated one-to-one with employment observed in that occupation, indicating that BG has been consistently representative of U.S. employment.

that the coefficient of interest is almost unchanged when including time fixed effects (-0.218, s.e.=0.100), so that our findings cannot be explained by an increasing share of low-skilled jobs being advertised online. Appendix Table 14 expands on this theme by estimating skill bias and broadening separately for two sub-samples (2010-2015 and 2016-2019), with almost identical results in each case.

In column 4, we introduce technology bigram fixed effects and now find a larger negative slope (0.493, s.e.=0.036), but also a larger constant term (63.898, s.e.=0.840). Taken at face value, the two estimates imply that new technologies take 68.08 years to reach the average level of college education among the U.S. workforce (30.3% in the 2015 ACS). In other words, the skill bias of a given new technology on average takes several generations to dissipate.

Panel B of Table 7 repeats this estimation using alternative measures of skill. It shows that, in the year of emergence, jobs in a new technology on average require 15.5 years of schooling, 22.6% of them require a post-graduate degree, and they pay an average wage of $75,521 (measured in 2015 dollars). All three skill indicators again decay significantly over time, at rates that would imply 77.6, 78.0, and 69.7 years to reach the average years of schooling, rate of post-graduate education, and wage of the U.S. population reported in the ACS.

All of these variations show (1) that job postings mentioning new technologies are strongly high-skill biased initially and (2) this skill bias decays significantly over time, albeit at a relatively slow rate, so that the skill bias of jobs associated with new technologies persists for multiple decades.

Both findings intersect with important branches of the literature studying the relationship between technology and inequality. First, they show direct evidence of the high-skill bias of new technologies, adding to a large literature that infers this skill bias from observed wage premia (e.g., Katz and Murphy, 1992). The findings suggest in a dramatic way that new technologies contribute to persistent inequalities between high- and low-skilled workers and, because pioneer locations of technologies are highly concentrated, also engender persistent inequalities across space. In this sense, innovation has a profound effect on regional disparities in economic opportunity.

Second, our finding of skill broadening provides direct evidence for a key assumption in the literature on automation: that the comparative advantage of high-skill workers in a new task erodes as the technology matures, pulling lower-skilled workers into working with a new technology over time. It is this key assumption that leads to balance in the “race between man and machine” in

Acemoglu and Restrepo (2018) and the related literature. Our evidence suggests this skill broadening indeed occurs in the data.

a. Mechanisms

Given these results, a key question is why skill broadening occurs in practice. The literature has suggested at least two, possibly complementary, channels. The first is standardization of new technologies – where research and customization become less important as new technologies mature and become standardized. That is, the new technology evolves over time into a standardized form that can more readily be used by less educated workers (Acemoglu et al., 2012; Acemoglu and Restrepo, 2018). The second is training or experience – over time, less educated workers may acquire training or experience that allows them to use new technologies, even if they do not have high levels of formal education (Nelson and Phelps, 1966; Galor and Moav, 2000).

Again, analyzing the context of the mention of the new technology within a given job posting can shed some light on these mechanisms. To this end, we use our keyword-based approach to systematically flag those job postings that mention a given new technology in conjunction with a requirement of training or experience with that technology (starting with seed phrases "training in," "knowledge of," "experience with," "familiar with," "knowhow of," and "proficiency in"). We again use the same iterative procedure combining our embedding vector model with human reading to settle on a list of keywords (shown in Appendix Table 15).

Appendix Figure 7 shows the proportion of RDP jobs declines significantly over time, so that more mature technologies have a lower share of jobs involved with RDP. At the same time, these RDP technology jobs skew heavily on the side of higher college requirements. At the same time, training / experience requirements with the new technology are positively, not negatively, associated with college requirements, so that training in a new technology and formal education appear to be complements, not substitutes in our data (Appendix Figure 8).

To assess to what extent these two channels can account for new technologies' skill broadening over time, Table 8 separately adds both as controls, to assess to what extent their inclusion can attenuate the estimated coefficient, β_{SB}. Column 1 reproduces our estimate from column 2 of Panel A, Table 7 for comparison (-0.288, s.e.=0.079). Column 4 shows that controlling for the inverse hyperbolic sine of the share of RDP jobs in the same technology attenuates this estimate by about 22% to -0.224 (s.e.=0.055). Columns 2 and 3 shows similar, albeit somewhat smaller, attenuations

when controlling separately for the share of R&D job postings and the share of job postings relating to production.[45] We conclude that technologies' transition from a focus on RDP towards a focus on use can account for part of the skill broadening we observe in the data.

By contrast, column 5 shows that controlling for the share of that technology's jobs that require training or experience in the technology results in no attenuation whatsoever (in fact, an increase) of our estimate of β_{SB}. In this sense, changes in training and experience in the technology cannot account for the pattern of skill broadening observed in the data.

We tentatively conclude that training and experience does not appear to be a substitute for formal education when it comes to required qualifications for jobs in new technologies, as measured in job postings. Instead, some of the observed skill broadening can indeed be accounted for by standardization of the technology over time.

6. Diffusion across Occupations, Industries, and Firms

Finally, before exploring the robustness of our main findings, we highlight the power of the data that we have developed to also characterize the spread of new technologies across other dimensions.

To assess the rate at which new technologies spread across occupations, firms, and industries, we extend the definition of $Normalized\ share_{c,\tau,t}$ to NAICS four-digit industries, SOC six-digit occupations, and firms for each technology (τ) and time (t), calculating the normalized share of job postings in each industry, occupation, and firm that mention a given new technology.[46] We then measure the coefficient of variation of $Normalized\ share_{c,\tau,t}$ across the segments.

Because the number of firms posting job advertisements online expands over time, we stratify our firm-technology-year sample by including only firms that post at least one job in each of our sample-years, before calculating the coefficient of variation.[47] This step focuses attention on

[45] For the sub-topic of research and development we start with the seed phrases "research and," "and develop," "and development," and "customization of" – a subset of our RDP seed keywords above – and proceed in the same manner. The remainder of the RDP keywords constitute the "produce" category.

[46] While the former two variables are included in the BG data (in each case, we use the finest level of disaggregation available from BG), the latter relies on our own matching algorithm described in Section 2.

[47] Hershbein and Kahn (2018) discuss this fact in some detail. The general increase in coverage of the BG data over time should not affect any of our main results. We discuss robustness to various weighting schemes in detail in Section 7.

10,496 larger firms, which on average post 865 job postings per year, effectively excluding variation coming from small and medium-sized businesses.

Spread across firms, occupations, and industries. Table 9, Panel A shows the results of a regression of the coefficient of variation calculated for each technology (τ) and time (t) on the year since emergence. Column 4 shows our already established results for locations for comparison. We find that while there is a decline in concentration as measured by the coefficient of variation for all four segments, there is a relatively (and significantly) larger decline across locations and firms (columns 4 and 3) than across industries and occupations (columns 2 and 1). While the coefficient of variation declines on average by 1.8% and 1.6% per year for CBSAs and firms, respectively, the corresponding declines are 0.7% and 0.4% for occupations and industries, respectively.[48] In fact, in column 1, this rate of decline across industries is statistically indistinguishable from zero.

Advantages for pioneer firms and industries. Following our procedure for pioneer locations, we define pioneer industries and firms for each technology as those with the most assigned patents in the ten years after the technology's emergence year that collectively account for 50% of the matched patents in a given new technology.[49] In Panel B, we explore the initial hiring advantage of pioneer firms and industries by estimating specification (4) for these additional dimensions. The table shows that pioneering firms have a strong initial advantage in job postings, with a 2,093% higher normalized share of job postings in the year of emergence for pioneer firms. Over time, this advantage again degrades significantly, at a rate of 2.3% per year. Consistent with the results in Panel A, this rate of decline is statistically indistinguishable from zero for industries.

Taken together, this evidence suggests new technologies initially generate hiring that is highly localized by location, firm, and industry. Over time, this hiring disperses, particularly across locations and across firms. Looking in more depth on a within-firm basis at the dynamics around the location of innovation and job creation is a fertile avenue for future exploration.

[48] The decay rates across CBSAs are 0.016 (0.004), 0.011 (0.003), and 0.003 (0.002) higher than industries, occupations, and firms, respectively. These coefficients are statistically significant at the 1%, 1%, and 20% level, respectively.

[49] See Section 3 of the Data Appendix for details on how we match patents to large firms and industries – matching patents to occupations makes little sense, so that we do not calculate pioneer occupations – and Appendix Table 16 for some examples.

7. Robustness Checks and Extensions

Finally, we conduct a broad range of robustness exercises to assess to what extent judgments we have made could have affected our primary results: "concentration in the development of impactful technologies," "region broadening," "pioneer-location advantage," "skill broadening," and "differential region-broadening by skill level."

To this end, we first re-trace our four steps of data construction to reexamine each of the main decisions we made in this automated process. In each case, we alter one aspect of the process, re-create our entire dataset, and re-run our main analyses. Table 10 reports the main estimates of interest, where the first line of each panel reproduces the results of our baseline specification for comparison.

Influential patents (Step 1 in Section 2). When isolating new bigrams associated with influential innovations, we retained only those that appear in patents accumulating a total of at least 1,000 weighted citations. Having some such threshold is necessary to maintain computational feasibility (to avoid having to cross-reference 1.5 million novel bigrams to Wikipedia and our other text sources). However, Panel A of Table 10 shows our results are almost invariant to altering this threshold. The panel shows four variations, with cutoffs ranging from 1,250 to 2,000, each producing almost identical results.[50]

Phrase length (Step 1 in Section 2). Our methodology easily extends to including trigrams, in addition to bigrams in the analysis. Repeating our steps 1-4 for trigrams adds 328 technology trigrams. 262 of these simply add another phrase to the set of bigrams already associated with a given technology (Wikipedia title) in our data. Perhaps the only substantive additions are "real time communications" and "injection molding machine" (see Appendix Table 17).

Adding unigrams is slightly more complicated due to their sheer number (about 2 million pass the threshold of 1,000 cite-weighted patents, simply because unigrams are more frequent than bigrams). To keep the number of candidate unigrams manageable, we focus on those with more than 100 mentions in earnings calls. Doing so adds 200 new technology unigrams, 53 of which again simply add another phrase to the set of phrases associated with a given technology already

[50] The reason for this stability is apparent in Appendix Figure 10, which shows a strong correlation between the number of cite-weighted patents and job postings in which a technology is mentioned across all novel bigrams (i.e., including bigrams with few cite-weighted patents). That is, variations in our minimum citations cutoff will on average tend to remove technologies that have little traction in the labor market.

identified in our bigram-based analysis. Appendix Table 18 shows examples among the 147 remaining unigrams. Overall, as expected, the unigram-based approach appears significantly noisier, with some clear false positives ("billable," "internets") and names in the mix ("USPS"). Nevertheless, broadening our approach in this way also yields some substantive additions, including, for example, "mRNA" and "Bluetooth."

Re-running our analyses including these sets of unigrams and trigrams again has no material effects on our results.

Human audit (Step 2 in Section 2). Rather than relying fully on our Wikipedia filter to determine whether or not a novel and influential bigram describes a technology (as opposed to increasingly visible problems or management techniques), we also conducted a human audit, where team members read through each Wikipedia title – technology bigram pair and removed all of those where the match appeared erroneous (e.g. "OS-level virtualization" matched to "programs running") and those where either the Wikipedia title or the technology bigram did not describe a technology according to the team member's judgment (e.g. "adverse event"). Appendix Table 3 marks each of the economically impactful technologies dropped under this audit (altogether 63 of 276 technologies with $EC \geq 100$). Doing so again has a negligible effect on our estimates.

Emergence years (Step 3 in Section 2). Our baseline approach to defining a technology's emergence year requires that technologies are mentioned in at least 100 cite-weighted patents prior to their year of emergence. Two variations in Panel D loosen (one cite-weighted patent) and tighten (200 cite-weighted patents) this requirement. A third variation abandons this approach altogether and instead fixes the emergence year as the first year in which the technology reaches 50% of its maximum cite-weighted patents achieved by a technology bigram in our sample. All of these variations again have a negligible effect on our main results.

Note that each of these variations in the robustness checks above alters the list of new technologies we uncover in small ways. For example, "fracking" may only show up in our data if we explicitly allow for unigrams, in addition to bigrams. Similarly, requiring 2,000 rather than 1,000 cite-weighted patents before including a new bigram from patents in our first step of data construction will obviously shorten the list of new technologies we produce. Our measure of success is thus not to always produce the one true list of new technologies that arose in the past 40 years. Such an absolutely true list does not exist. Instead, the key is that our language-based approach produces a

list of technologies that is representative of new technologies in a statistical sense. The fact that all of the variations above produce very similar econometric results is evidence that we meet this bar.

Alternative weighting schemes (Step 4 in Section 2). Because the coefficient of variation, as well as other constructed moments at the technology-time level, become noisy with insufficient data, our baseline specifications down-weight technologies that are mentioned in relatively few job postings. Panel E repeats all analyses (i) with unweighted regressions, (ii) without the requirement of a minimum number of mentions in job postings, and (iii) with weights proportional to the natural logarithm of the number of job postings associated with the technology. We also re-run our entire analysis after collapsing technology bigrams at the technology (Wikipedia title) level. Again, none of these variations materially affect our results.

In the final line of the panel, we re-calculate our list of economically impactful technologies using an emergence-year-normalized number of earnings calls mentions: For each technology bigram *i* with year of emergence t_0, we divide the number of earnings calls appearances by the average number of earnings calls for all bigrams with year of emergence t_0. This adjustment alters our list of influential technologies by controlling for the different number of years that the various bigrams had to be mentioned in earnings calls. Again, doing so has little influence on our results.

Representativeness of the BG sample. To further address any concerns relating to the possibly changing composition of the BG data over time, Appendix Table 14 shows additional variations of Table 7, Panel A, column 2 where we (i) include the 2007 data and (ii) estimate our baseline coefficient separately for two sample periods (2010-2015 and 2016-2019).

Standard errors. We also explore the robustness of the results relative to the treatment of the standard errors. These examine again the four regressions that were analyzed in Table 10. We explore in Table 11 the impact on the standard errors of different clustering approaches: clustering the observations not by associated Wikipedia entries ("technologies"), but rather by the individual technology bigram, the year, and (in the case of the regression from Table 4) the CBSA, state, and the interaction between the CBSA and the associated Wikipedia entry. We also present bootstrapped standard errors, drawn from 1,000 replications with replacement. The changes have little effect on the significance of the results.

8. Conclusion

Policymakers in many parts of the world devote enormous energy to fostering nascent technologies, ranging from efforts to support academic research to luring start-ups from other cities and nations. Such infant industry strategies are often predicated on the notion that early advantages in innovation and employment will yield lasting benefits for regions, particularly in the form of high-quality employment.

Using the full text of patents, job postings, and earnings conference calls, we introduce in this paper an approach to understand which new technologies affect jobs and businesses and to trace their diffusion across regions, industries, occupations, and firms. We can then map the spread of new technologies in these dimensions, focusing on the hiring associated with each important innovation.

We highlight first that the locations where economically impactful technologies are developed are geographically highly concentrated, with a handful of urban areas contributing the bulk of the early patenting and early employment within influential new technologies. One striking figure is that 56% of the pioneering locations for the most economically impactful technologies are in two parts of the U.S. – Silicon Valley and the Northeast Corridor. Second, despite this initial concentration, jobs relating to new technologies spread out geographically. But this rate of diffusion is extremely slow, happening over several decades rather than in just a few years. Locally developed technologies continue to offer long-lasting benefits for jobs in their pioneer locations for multiple decades. Third, jobs relating to new technologies are highly skill biased – 57% of the initial jobs associated with a given new technology require a college degree. Over time, the mean required skill levels of the new jobs decline, albeit at a very slow pace. Fourth, low-skill jobs associated with the use of a given new technology spread out geographically significantly faster than high-skill ones, so that the pioneer locations where the technology was invented host a disproportionate share of high-skilled jobs relating to that new technology for several decades after its year of emergence.

Combined with the extreme spatial concentration of the most economically impactful innovations, this pioneer advantage engenders large and persistent regional disparities in economic opportunity, giving a handful of U.S. locations a lasting advantage in high-skill jobs.

Beyond these core results of our analysis, the development and spread of new technologies are key objects of interest in multiple fields of economics. As we suggest in Section 6, these techniques developed here should have applications for studies of firm-level technological adoption and implementation. More generally, we hope the text-to-data techniques we develop and data that we provide as part of this paper may prove useful in addressing a range of additional research questions in the study of economic growth, inequality, entrepreneurship, and firm dynamics.

References

Abis, Simona, and Laura Veldkamp. "The changing economics of knowledge production." Working paper 3570130, SSRN (2020).

Acemoglu, Daron, and David Autor. "Skills, tasks and technologies: Implications for employment and earnings." In Orley Ashenfelter and David Card (editors), Handbook of Labor Economics. New York, Elsevier, volume 4, chapter 12, pp. 1043-1171 (2011).

Acemoglu, Daron, Gino Gancia, and Fabrizio Zilibotti. "Competing engines of growth: Innovation and standardization." Journal of Economic Theory 147 (2012): 570-601.

Acemoglu, Daron, and Joshua Linn. "Market size in innovation: Theory and evidence from the pharmaceutical industry." Quarterly Journal of Economics 119 (2004): 1049-90.

Acemoglu, Daron, and Pascual Restrepo. "The race between man and machine: Implications of technology for growth, factor shares, and employment." American Economic Review 108 (2018): 1488-1542.

Acemoglu, Daron, and Pascual Restrepo. "Robots and jobs: Evidence from US labor markets." Journal of Political Economy 128 (2020): 2188-2244.

Agrawal, Ajay. Joshua Gans, and Avi Goldfarb (editors). The Economics of Artificial Intelligence: An Agenda. Chicago, University of Chicago Press (2019).

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. "The skill complementarity of broadband internet." Quarterly Journal of Economics 130 (2015): 1781–1824.

Arora, Ashish, Sharon Belenzon and Lia Sheer. "Knowledge spillovers and corporate investment in scientific research." American Economic Review 111 (2021) 871-898.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. "The evolution of work in the United States." American Economic Journal: Applied Economics 12 (2020): 1-34.

Autor, David H., Caroline Chin, Anna Salomons, and Bryan Seegmiller. "New frontiers: the origins and content of new work, 1940-2018." Quarterly Journal of Economics, qjae008 (2024).

Autor, David H., David Dorn, Gordon H. Hanson, Gary Pisano, and Pian Shu. "Foreign competition and domestic innovation: Evidence from US patents." American Economic Review: Insights 2 (2020): 357-374.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. "Trends in US wage inequality: Revising the revisionists." Review of Economics and Statistics 90 (2008): 300-323.

Autor, David H., Lawrence F. Katz, and Alan Krueger. “Computing inequality: Have computers changed the labor market?” Quarterly Journal of Economics 113 (1998): 1169-1213.

Autor, David H., Frank Levy, and Richard J. Murnane. “The skill content of recent technological change: An empirical exploration.” Quarterly Journal of Economics 118 (2003): 1279-1334.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. "Measuring economic policy uncertainty." Quarterly Journal of Economics 131 (2016): 1593-1636.

Bekkerman, Ron, and James Allan. “Using bigrams in text categorization.” Technical report IR-408, Center of Intelligent Information Retrieval, University of Massachusetts at Amherst (2004).

Berman, Eli, John Bound, and Zvi Griliches. “Changes in the demand for skilled labor within U.S. manufacturing: Evidence from the Annual Survey of Manufacturers.” Quarterly Journal of Economics 109 (1994): 367-397.

Bloom, Nicholas, Rafaella Sadun, and John Van Reenen. “Management as a technology?” Working paper no. 22327, National Bureau of Economic Research (2016).

Buera, Francisco J., and Ezra Oberfield. “The global diffusion of ideas.” Econometrica 88 (2020): 83-114.

Bushee, Brian J., Dawn A. Matsumoto, and Gregory S. Miller. "Open versus closed conference calls: The determinants and effects of broadening access to disclosure." Journal of Accounting and Economics 34 (2003): 149-180.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. “The structure of economic news.” Working paper no. 26648, National Bureau of Economic Research (2020).

Caprettini, Bruno, and Hans-Joachim Voth. “Rage against the machines: Labor-saving technology and unrest in industrializing England.” American Economic Review: Insights 2 (2020): 305-320.

Caselli, Francesco. "Technological revolutions." American Economic Review 89 (1999): 78-102.

Comin, Diego, and Bart Hobijn. “Cross-country technology adoption: Making the theories face the facts.” Journal of Monetary Economics 51 (2004): 39-83.

Comin, Diego, and Bart Hobijn. “An exploration of technology diffusion.” American Economic Review 100 (2010): 2031–59.

Davies, Mark. "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights." International Journal of Corpus Linguistics 14 (2009): 159-190.

Deming, David J., and Kadeem Noray. "Earnings dynamics, changing job skills, and STEM careers." Quarterly Journal of Economics. 135 (2020): 1965-2005.

Flynn, Joel P., and Karthik A. Sastry. “The macroeconomics of narratives,” Working paper 4140751, SSRN (2022).

Forman, Chris, Avi Goldfarb, and Shane Greenstein, "Agglomeration of invention in the Bay Area: Not just ICT." American Economic Review Papers and Proceedings 106 (2016): 146-151.

Furman, Jeffrey L., and Megan J. MacGarvie. “Academic science and the birth of industrial research laboratories in the U.S. pharmaceutical industry.” Journal of Economic Behavior and Organization 63 (2007): 756-776.

Galor, Oded, and Omer Moav. "Ability-biased technological transition, wage inequality, and economic growth." Quarterly Journal of Economics 115 (2000): 469-497.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. “Text as data.” Journal of Economic Literature 57 (2019): 535-574.

Glaeser, Edward L., Sari P. Kerr, and William R. Kerr. "Entrepreneurship and urban growth: An empirical assessment with historical mines." Review of Economics and Statistics 97 (2015): 498-520.

Goldin, Claudia D., and Lawrence F. Katz. “The origins of technology-skill complementarity.” Quarterly Journal of Economics 113 (1998): 683-732.

Goldin, Claudia D., and Lawrence F. Katz. The Race Between Education and Technology. Cambridge, Harvard University Press (2008).

Gompers, Paul, Josh Lerner, and David Scharfstein. "Entrepreneurial spawning: Public corporations and the genesis of new ventures, 1986 to 1999." Journal of Finance 60 (2005): 577-614.

Gordon, Robert J. The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War. Princeton, Princeton University Press (2016).

Greenstone, Michael, Richard Hornbeck and Enrico Moretti, "Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings." Journal of Political Economy 118 (2010): 536-598.

Griliches, Zvi, "Hybrid corn: An exploration in the economics of technological change." Econometrica 25 (1957): 501–22.

Hall, Bronwyn H., Jacques Mairesse, and Laure Turner. "Identifying age, cohort, and period effects in scientific research productivity: Discussion and illustration using simulated and actual data on French physicists." Economics of Innovation and New Technologies 16 (2007): 159-177.

Handley, Kyle, and J.F. Li. "Measuring the effects of firm uncertainty on economic activity: New evidence from one million documents." Working paper no. 27896, National Bureau of Economic Research (2020).

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. "Firm-level political risk: Measurement and effects." Quarterly Journal of Economics 134 (2019): 2135–2202.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. "Firm-level exposure to epidemic diseases: Covid-19, SARS, and H1N1." Working paper no. 26971, National Bureau of Economic Research (2021).

Hershbein, Brad, and Lisa B. Kahn. "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." American Economic Review 108 (2018): 1737-72.

Hoberg, Gerard, and Gordon Phillips. "Text-based network industries and endogenous product differentiation." Journal of Political Economy 124 (2016): 1423-65.

Jaffe, Adam B. "Real effects of academic research." American Economic Review 79 (1989): 957–970.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic localization of knowledge spillovers as evidenced by patent citations." Quarterly Journal of Economics 108 (1993): 577–598.

Jewkes, John, David Sawers, and Richard Stillerman. The Sources of Invention. New York, St. Martin's Press (1969).

Katz, Lawrence F., and Kevin M. Murphy. "Changes in relative wages, 1963-1987: Supply and demand factors." Quarterly Journal of Economics 107 (1992): 35-78.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. "Measuring technological innovation over the long run." American Economic Review: Insights 3 (2021): 303-320.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar. "Who profits from patents? Rent-sharing at innovative firms." Quarterly Journal of Economics 134 (2019): 1343-1404.

Kogan, Leonid, Dimitris Papanikolaou, Lawrence D. Schmidt, and Bryan Seegmiller. "Technology, vintage-specific human capital, and labor displacement: Evidence from linking patents with occupations." Working paper no. 29552, National Bureau of Economic Research (2022).

Krueger, Alan B. "How computers have changed the wage structure: Evidence from microdata, 1984-1989." Quarterly Journal of Economics 108 (1993): 33-60.

Krusell, Per, Lee E. Ohanian, José-Víctor Ríos-Rull, and Giovanni L. Violante. "Capital-skill complementarity and inequality: A macroeconomic analysis." Econometrica 68 (2000): 1029-53.

Lanjouw, Jean O., Ariel Pakes, and Jonathan Putnam. "How to count patents and value intellectual property: The uses of patent renewal and application data." Journal of Industrial Economics 46 (1998): 405-432.

Lerner, Josh, and Amit Seru. "The use and misuse of patent data: Issues for finance and beyond." Review of Financial Studies 35 (2022): 2667–2704.

Lind, Nelson, and Natalia Ramondo. "Global innovation and knowledge diffusion." Working paper no. 29629, National Bureau of Economic Research (2022).

Matsumoto, Dawn, Maarten Pronk, and Erik Roelofsen. "What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions." Accounting Review 86 (2011): 1383-1414.

Merton, Robert K. "The Matthew effect in science: The reward and communication systems of science are considered." Science 159 (3810) (1968): 56-63.

Michaels, Guy, Ashwini Natraj, and John Van Reenen. “Has ICT polarized skill demand? Evidence from eleven countries over 25 years.” Review of Economics and Statistics 96 (2014): 60-77.

Mokyr, Joel. The Lever of Riches: Technological Creativity and Economic Progress. New York, Oxford University Press (1992).

Moretti, Enrico. "The effect of high-tech clusters on the productivity of top inventors." American Economic Review 111 (2021): 3328-75.

Moscona, Jacob. "Environmental catastrophe and the direction of invention: Evidence from the American Dust Bowl." Unpublished working paper, Massachusetts Institute of Technology (2020).

Moser, Petra, Alessandra Voena, and Fabian Waldinger. “German Jewish émigrés and US invention.” American Economic Review 104 (2014): 3222-55.

Nelson, Richard R., and Edmund S. Phelps. "Investment in humans, technological diffusion, and economic growth." American Economic Review 56 (1966): 69-75.

Organisation for Economic Cooperation and Development. The Measurement of Scientific and Technological Activities: Proposed Guidelines for Collecting and Interpreting Technological Innovation Data (third edition). Paris, OECD, Chapter 3 (2005).

Piketty, Thomas, and Emmanuel Saez. "Income inequality in the United States, 1913–1998." Quarterly Journal of Economics 118 (2003): 1-41.

Popp, David. "Induced innovation and energy prices." American Economic Review 92 (2002): 160-180.

Rogers, Everett M., Diffusion of Innovations. New York, Free Press (1962).

Sautner, Zacharias, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang. “Firm-level climate change exposure.” Journal of Finance 78 (2023): 1449-98.

Schmookler, Jacob. Invention and Economic Growth. Cambridge, Harvard University Press (1966).

Schumpeter, Joseph A. Capitalism, Socialism, and Democracy. New York, Harper (1942).

Song, Jae, David J. Price, Faith Guvenen, Nicholas Bloom, and Till Von Wachter. "Firming up inequality." Quarterly Journal of Economics 134 (2019): 1-50.

Squicciarini, Mara P., and Nico Voigtländer. "Human capital and industrialization: Evidence from the age of enlightenment." Quarterly Journal of Economics 130 (2015): 1825-83.

Syverson, Chad. "What determines productivity?" Journal of Economic Literature 49 (2011): 326-365.

Tambe, Prasanna. "Big data investment, skills, and firm value." Management Science 60 (2014): 1452-69.

Tambe, Prasanna, and Lorin M. Hitt. "Now IT's personal: Offshoring and the shifting skill composition of the US information technology workforce." Management Science 58 (2012): 678-695.

Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson. "Digital capital and superstar firms." Working paper no. 28285, National Bureau of Economic Research (2020).

Tan, Chade-Meng, Yuan-Fang Wang, and Chan-Do Lee. "The use of bigrams to enhance text categorization." Information Processing & Management 38 (2002): 529–546.

Tyson Laura D., and Michael Spence. "Exploring the effects of technology on income and wealth inequality." In Heather Boushey, J. Bradford DeLong, and Marshall Steinbaum (editors), After Piketty: The Agenda for Economics and Inequality. Cambridge, Harvard University Press, pp. 170–208 (2017).

United States Patent and Trademark Office. Performance and Accountability Report. Washington, USPTO (2020).

Van Reenen, John. "The creation and capture of rents: Wages and innovation in a panel of U. K. companies." Quarterly Journal of Economics 111 (1996): 195-226.

Vance, J.D. "One on one interview with Ohio US Senate candidate JD Vance," https://www.youtube.com/watch?v=4FIapZ88BJQ (October 19, 2022).

Webb, Michael. "The impact of artificial intelligence on the labor market." Unpublished working paper, Stanford University (2020).

Zucker, Lynne, Michael Darby, and Marilynn B. Brewer. “Intellectual human capital and the birth of U.S. biotechnology enterprises.” American Economic Review 88 (1998): 290-306.

Table 1 – Top technologies by year of emergence

Emergence year	Wikipedia title (technology)	Technology bigrams	Number of job postings
1979	Hard disk drive	hard disk; disk drive	34,211
1980	Barcode reader	barcode reader; code reader; code scanner; barcode scanner	43,279
1981	Laser diode	emitting laser; diode laser; semiconductor laser; laser diode	7,284
1982	Personal computer	personal computer	1,752,726
1983	Flat-panel display	panel display; flat panel	27,369
1984	User interface	user interface	747,586
1985	Mobile phone	mobile telephone; cellular telephone; phones mobile; cellular phone; mobile phone; cell phone	1,832,787
1986	Facial recognition system	frt system; recognition software; recognition system; recognition technology; facial recognition	25,109
1987	Digital video	digital video	88,887
1988	Model organism	animal model	24,722
1989	Mobile device	held computer; computer device; handheld computer; mobile device	1,046,079
1990	Debit card	cards debit; card debit; debit card	260,282
1991	Flash memory	flash device; nand flash; flash memory	22,882
1992	Machine learning	learning algorithm; machine learning	491,252
1993	Financial instrument	financial instrument	43,944
1994	Active users	active user	39,671
1995	Hybrid electric vehicle	hybrid electric	8,207
1996	Digital content	digital content	144,775
1997	Multicore processor	multi core; core processor	29,643
1998	Information privacy	data protection	176,110
1999	Unmanned aerial vehicle	aerial vehicle; unmanned aerial	24,148
2000	Transaction account	transaction account	13,012
2001	Smartphone	smart phone	910,856
2002	Online game	online game	15,254
2003	Social networking service	networking site; social networking	244,610
2004	Electronic discovery	electronic format	56,438
2005	LED circuit	led driver	2,575
2006	Augmented reality	augmented reality	20,537
2007	Self-driving car	autonomous vehicle	18,641

Notes: This table reports the top technology by number of mentions in earnings calls (in column 2) for every year of emergence between 1976 and 2007 (in column 1). Column 3 lists the associated technology bigram(s). Column 4 lists the number of job postings that the bigram appears in. For the year of emergence 1999, the most frequent technology in earnings calls was "adverse event." We replace "adverse event" (as it gets dropped in our human audit) with the next most frequent technology, "unmanned aerial vehicle." Column 4 reports the number of job postings associated with the technology. See Section 2.c of the main text for details.

Table 2 – Examples of technologies and pioneer locations

Machine Learning (1992)

CBSA	State	Pct. Patents
New York-Newark-Jersey City	NY-NJ-PA	24%
Seattle-Tacoma-Bellevue	WA	13%
San Jose-Sunnyvale-Santa Clara	CA	12%
San Francisco-Oakland-Hayward	CA	9%

Digital Imaging (1992)

CBSA	State	Pct. Patents
Rochester	NY	18%
San Jose-Sunnyvale-Santa Clara	CA	12%
San Francisco-Oakland-Hayward	CA	7%
Fort Collins	CO	6%
Greeley	CO	5%
Worcester	MA-CT	4%

Hybrid Electric (1995)

CBSA	State	Pct. Patents
Detroit-Warren-Dearborn	MI	33%
Ann Arbor	MI	10%
Indianapolis-Carmel-Anderson	IN	8%

Smart Phone (2001)

CBSA	State	Pct. Patents
San Francisco-Oakland-Hayward	CA	18%
San Jose-Sunnyvale-Santa Clara	CA	18%
Seattle-Tacoma-Bellevue	WA	6%
New York-Newark-Jersey City	NY-NJ-PA	5%
Los Angeles-Long Beach-Anaheim	CA	4%

Notes: The table shows pioneer CBSAs (in column 1), along with their state (in column 2) and the percentage of early cite-weighted patents accounted for by these CBSAs (in column 3) for a sample of four example technology bigrams – "machine learning," "digital imaging," "hybrid electric," and "smart phone." Early patents are defined as patents filed within ten years of the emergence year of technology. Each technology bigram's emergence year is given in parentheses. See Section 2.c of the main text for details.

Table 3 – Geographic concentration of patents, skill, and employment

	Total Number	Share Top 5 CBSAs	Top 5 CBSAs
	(1)	(2)	(3)
Panel A: Geographic concentration of U.S. patents			
Economically impactful	1,044,351	42.1%	**San Jose-Sunnyvale-Santa Clara CA** **San Francisco-Oakland-Hayward, CA** **New York-Newark-Jersey City, NY-NJ-PA** **Seattle-Tacoma-Bellevue, WA** **Boston-Cambridge-Newton, MA-NH**
All New Technologies	1,623,800	33.3%	**San Jose-Sunnyvale-Santa Clara, CA** **San Francisco-Oakland-Hayward, CA** **New York-Newark-Jersey City, NY-NJ-PA** Los Angeles-Long Beach-Anaheim, CA **Boston-Cambridge-Newton, MA-NH**
All Patents	3,146,114	32.4%	**San Jose-Sunnyvale-Santa Clara CA** **New York-Newark-Jersey City, NY-NJ-PA** **San Francisco-Oakland-Hayward, CA** Los Angeles-Long Beach-Anaheim, CA Chicago-Naperville-Elgin, IL-IN-WI
Most Cited	1,044,351	32.7%	**San Jose-Sunnyvale-Santa Clara CA** **San Francisco-Oakland-Hayward, CA** **New York-Newark-Jersey City, NY-NJ-PA** Los Angeles-Long Beach-Anaheim, CA Chicago-Naperville-Elgin, IL-IN-WI
Panel B: Geographic concentration of skill and employment			
College Graduates	51.5 million	22.5%	**New York-Newark-Jersey City, NY-NJ-PA** Los Angeles-Long Beach-Anaheim, CA Chicago-Naperville-Elgin, IL-IN-WI Washington-Arlington-Alexandria, DC-VA-MD-WV **San Francisco-Oakland-Hayward, CA**
Employed	156.5 million	18.9%	**New York-Newark-Jersey City, NY-NJ-PA** Los Angeles-Long Beach-Anaheim, CA Chicago-Naperville-Elgin, IL-IN-WI Dallas-Fort Worth-Arlington, TX Houston-The Woodlands-Sugar Land, TX

Notes: This table reports the concentration of patents, skill, and employment across CBSAs in the U.S. The measures of skill and employment are obtained from the 2015 American Communities Survey. A patent is considered an economically impactful/new technology patent if it mentions at least one bigram associated with an economically impactful/new technology more than once. The row "Most Cited" shows the geographic concentration of the 1,044,351 patents with the most normalized citations for comparison. This number is chosen to equal the number of patents mentioning a economically impactful technology. CBSAs in **bold** are those in the top five for patents which mention economically impactful technologies.

Table 4 – Region broadening

Panel A: Main specifications			
	$Coefficient\ of\ Variation_{\tau,t}$		
Sample	EC>=100	All	
	(1)	(2)	(3)
$Years\ since\ emergence_{\tau,t}$	-0.068***	-0.065***	-0.153***
	(0.026)	(0.024)	(0.012)
Constant (CV at t=$t_{\tau,0}$)	5.577***	6.212***	8.269***
	(0.645)	(0.585)	(0.271)
R-squared	0.019	0.013	0.825
N	4,270	8,347	8,347
Bigrams	428	835	835
Bigram FE	NO	NO	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title
Years to zero CV	82.12	95.52	54.07
Panel B: Alternative measures of geographic concentration			
	$\frac{N_{c,\tau,t}^{Top-5}}{N_{c,\tau,t}^{All}}$	$Pct.(N_{c,\tau,t} \leq 0.1)$	$\sum_i (N_{c,\tau,t}-1)^2$
	(1)	(2)	(3)
$Years\ since\ emergence_{\tau,t}$	-1.724***	-1.117***	-173.965**
	(0.136)	(0.088)	(77.247)
Cons (concentration at t=$t_{\tau,0}$)	88.781***	95.514***	13,650.448***
	(3.181)	(2.063)	(1,808.008)
R-squared	0.846	0.922	0.679
N	8,347	8,347	8,347
Bigrams	835	835	835
Bigram FE	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title
Years to zero CV	51.49	85.48	78.47

Notes: This table reports the results from regressions at the technology bigram x year level. The dependent variable is a measure of the geographic concentration of a given technology bigram's job postings in a given year. The independent variable – years since emergence – is the number of years that have elapsed since the technology's year of emergence. Panel A reports results using our baseline measure of geographic concentration – the coefficient of variation of the normalized share of a technology bigram's job postings across CBSAs. Panel B reports results using three alternative measures of geographic concentration – the mean normalized share of a technology's job postings in the top five CBSAs relative to the mean normalized share across all CBSAs, the percentage of CBSAs with a normalized share of a technology's job postings of less than 10% (that is, the representation of CBSAs with almost no activity associated with that bigram), and the sum of squared deviations of the normalized share from one (similar to the Herfindahl-Hirschman Index). Column 1 of Panel A is restricted to the sample of technology bigrams that appear in at least 100 earnings calls. The other regressions use all technology bigrams that appear in at least 1000 job postings in our sample. Observations are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. The normalized share of job postings is capped at the 99th percentile of non-zero observations. Standard errors are clustered by Wikipedia title (technology). All specifications indicate fixed effects used. Years to zero CV are calculated by dividing the constant by the coefficient estimate on the years since emergence.

Table 5 – Pioneer location advantage in technology hiring

	$Normalized\ Share_{c,\tau,t}$			
Sample:	$EC_\tau \geq 100$		All	
	(1)	(2)	(3)	(4)
$Pioneer_{c,\tau}$	0.311***	1.084***	1.321***	1.282***
	(0.076)	(0.309)	(0.254)	(0.243)
$Pioneer_{c,\tau} * Years\ since\ emg_{\tau,t}$		-0.032**	-0.035***	-0.034***
		(0.013)	(0.011)	(0.010)
$Pioneer\ Neighbor_{c,\tau}$				0.158***
				(0.057)
$Pioneer\ Neighbor_{c,\tau} * Years\ since\ emg_{\tau,t}$				-0.004
				(0.003)
R-squared	0.038	0.038	0.030	0.030
N	3,965,122	3,965,122	7,751,024	7,751,024
Bigrams	428	428	835	835
Bigram FE	YES	YES	YES	YES
CBSA FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title
$Rate\ of\ decline\ per\ year$		-0.029	-0.027	-0.026
		(0.005)	(0.003)	(0.003)
Implied years to zero advantage		**33.88**	**37.74**	**38.26**

Notes: This table reports results from regressions of the $Normalized\ Share_{c,\tau,t}$ (for each CBSA x technology bigram x year) on a dummy indicating the pioneer status of the CBSA and the interaction of this dummy with the number of years that have elapsed since the bigram's emergence. The dummy variable $Pioneer\ Neighbor_{c,\tau}$ takes value one for non-pioneer CBSAs that are within 100 miles of the technology's pioneer locations. Columns 1 and 2 are restricted to the sample of technology bigrams that appear in at least 100 earnings calls. The other regressions use all technology bigrams that appear in at least 1000 job postings in our sample. Observations are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. The normalized share of job postings is capped at the 99th percentile of non-zero observations. All specifications indicate fixed effects used. Standard errors are clustered by Wikipedia title (technology). The rate of decline per year is calculated as $\frac{\beta_D}{\beta_P}$, where β_P is the coefficient on $Pioneer_{c,\tau}$ and β_D is the coefficient of $Pioneer_{c,\tau} * Years\ since\ emg_{\tau,t}$.

Table 6 – Mechanisms: Spread of high vs. low-skill jobs;
Spread of research, development, and production jobs vs. use jobs

Panel A: Region-broadening regressions				
	$\log(Coefficient\ of\ Variation)_{\tau,t}$			
	All			
Sample:	(1) High-Skill Job Postings	(2) Low-Skill Job Postings	(3) RDP Job Postings	(4) Use Job Postings
$Years\ since\ emergence_{\tau,t}$	-0.027***	-0.038***	-0.014***	-0.036***
	(0.002)	(0.003)	(0.002)	(0.002)
R-squared	0.837	0.845	0.736	0.883
N	8,069	8,069	6,033	6,033
Bigram FE	YES	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title
Panel B: Pioneer advantage regressions				
	$Normalized\ Share_{c,\tau,t}$			
	All			
Sample:	(1) High-Skill Job Postings	(2) Low-Skill Job Postings	(3) RDP Job Postings	(4) Use Job Postings
$Pioneer_{c,\tau}$	1.319***	1.127***	1.974***	1.581***
	(0.233)	(0.255)	(0.595)	(0.301)
$Pioneer_{c,\tau} * Years\ since\ emg_{\tau,t}$	-0.029***	-0.036***	-0.021	-0.030**
	(0.010)	(0.010)	(0.027)	(0.012)
R-squared	0.016	0.012	0.003	0.020
N	8,581,946	8,395,144	5,618,723	7,769,596
Bigram FE	YES	YES	814	837
CBSA FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	YES	YES
$Rate\ of\ decline\ per\ year$	-0.022	-0.032	-0.011	-0.019
	0.004	0.003	0.011	0.005
Implied years to zero advantage	45.48	31.31	104.28	51.76

Notes: This table reports region-broadening regressions at the technology bigram x year level (Panel A) and pioneer advantage regressions at the technology bigram x year x CBSA level (Panel B). Columns 1 and 2 show separate regressions for high-skill (column 1) and low-skill (column 2) job postings. Column 3 shows regressions for research, development, and production-related job postings (RDP); column 4 for job postings relating to the use of the technology. For definitions of these concepts, see Section 4.c of the main text. All specifications use technology bigrams that appear in at least 100 earnings calls. Observations are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. All specifications indicate fixed effects used. Standard errors are clustered by Wikipedia title (technology). Panel A reports results from regressions of $log(Coefficient\ of\ Variation)_{\tau,t}$ on the number of years since the technology's year of emergence. In a stacked specification, the difference between coefficient on $Years\ since\ emergence_{\tau,t}$ in columns 1 and 2 is -0.011 (S.E. = 0.003, p-val. = 0.001). The difference between the coefficient on $Years\ since\ emergence_{\tau,t}$ in columns 3 and 4 is -0.023 (S.E. = 0.003, p-val = 0.000). Both differences are thus statistically distinguishable from zero. Panel B reports results from regressions of the $Normalized\ share_{c,\tau,t}$ (for each CBSA, bigram, and year) on a dummy indicating pioneer status of the CBSA and on the interaction of this dummy with the number of years that have elapsed since bigram's emergence. The normalized share of job postings is capped at the 99th percentile of non-zero observations. In a stacked specification, the difference between estimates of rate of decline per year in columns 1 and 2 is 0.010 (S.E. = 0.005, p-val = 0.026). Similarly, the difference between the rate of decline per year in columns 3 and 4 is 0.007 (S.E. = 0.007, p-val = 0.518). The rate of decline per year is calculated as $\frac{\beta_D}{\beta_P}$, where β_P is the coefficient on $Pioneer_{c,\tau}$ and β_D is the coefficient of $Pioneer_{c,\tau} * Years\ since\ emg_{\tau,t}$.

Table 7 – Skill broadening

Panel A: Main specifications				
	$Share\ College\ Educated_{\tau,t} * 100$			
Sample	EC >= 100	All		
	(1)	(2)	(3)	(4)
Constant (Sh. Col. Ed. at t=$t_{\tau,0}$)	57.078***	59.095***	57.475***	63.898***
	(2.135)	(1.794)	(2.294)	(0.840)
$Years\ since\ emergence_{\tau,t}$	-0.228**	-0.288***	-0.218***	-0.493***
	(0.092)	(0.079)	(0.100)	(0.036)
R-squared	0.017	0.019	0.024	0.910
N	4,270	8,347	8,347	8,347
Bigrams	428	835	835	835
Year FE	NO	NO	YES	NO
Bigram FE	NO	NO	NO	YES
Standard Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title
Implied years to average skill	**117.23**	**100.03**	**124.37**	**68.08**

Panel B: Alternative measures of skill			
	(1)	(2)	(3)
	$Years\ of\ Schooling_{\tau,t}$	$Share\ Post\ Graduates_{\tau,t} * 100$	$Average\ Wage_{\tau,t}$
Constant (Skill at t=$t_{\tau,0}$)	15.504***	22.617***	75,521.317***
	(0.047)	(0.456)	(840.562)
$Years\ since\ emergence_{\tau,t}$	-0.024***	-0.149***	-505.134***
	(0.002)	(0.020)	(35.986)
R-squared	0.915	0.905	0.889
N	8,347	8,347	8,347
Bigrams	835	835	835
Bigram FE	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title
Implied years to average skill	**77.59**	**78.03**	**69.72**

Notes: This table reports results from regressions at the technology bigram x year level. The dependent variable is a measure of the average skill requirement of a technology bigram's job postings in a given year. The independent variable is the number of years that have elapsed since the technology's emergence. The dependent variable in Panel A is the average share of job postings mentioning technology bigram τ in year t that require a college degree. Panel B shows results corresponding to column 4 of Panel A for alternative measures of skill associated with technology bigram job postings: average years of schooling (column 1), share of post-graduates (in column 2), and average wage (in column 3). Column 1 of Panel A is restricted to the sample of technology bigrams that appear in at least 100 earnings calls. The other regressions use all technology bigrams that appear in at least 1000 job postings in our sample. Observations are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. All specifications indicate fixed effects used. Standard errors are clustered by Wikipedia title (technology). The row "Implied years to average skill" is determined by $-(Constant - Average\ Population\ Skill)/\beta_{SB}(Years\ since\ emergence_{\tau,t})$, where $Average\ Population\ Skill$ represents the weighted average skill of the US population according to the 2015 ACS Survey.

Table 8 – Skill broadening mechanisms: Research, development, and production and training jobs

Sample	$Share\ College\ Educated_{\tau,t}$				
	All				
	(1)	(2)	(3)	(4)	(5)
$Years\ since\ emergence_{\tau,t}$	--0.288***	--0.231***	--0.268***	--0.224***	--0.340***
	(0.079)	(0.056)	(0.063)	(0.055)	-0.069
$Share\ of\ R\&D\ Postings_{\tau,t}$, IHS		6.938***			
		(0.366)			
$Share\ of\ Produce\ Postings_{\tau,t}$, IHS			5.528***		
			(0.388)		
$Share\ of\ RDP\ Postings_{\tau,t}$, IHS				6.795***	
				(0.381)	
$Share\ Training\ Required_{\tau,t}$, IHS					5.553***
					-0.459
$Constant$	59.095***	43.482***	46.661***	39.136***	39.619***
	(1.794)	(1.499)	(1.670)	(1.662)	(2.413)
R-squared	0.019	0.432	0.276	0.407	0.236
N	8,347	8,347	8,347	8,347	8,347
Standard Errors (Cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title	Wiki Title

Notes: This table reports results from regressions at the technology bigram x year level. Column 1 replicates the specification in Table 7, Panel A, column 2. Columns 2-5 add additional controls: the inverse hyperbolic sine (IHS) of the share of the technology's job postings relating to research and development (column 2), the share of the technology's job postings relating to the technology's production (column 3), the share of the technology's job postings relating to research, development, and production (column 4), and the share of the technology's job postings requiring training in the technology (column 5). Observations are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. Standard errors are clustered by Wikipedia title (technology).

Table 9 – Broadening and pioneer advantage across different dimensions

Panel A: Broadening				
	$Coefficient\ of\ Variation_{\tau,t}$			
	(1) Industries	(2) Occupations	(3) Firms	(4) CBSAs
$Years\ since\ emergence_{\tau,t}$	-0.018	-0.056***	-0.354***	-0.153***
	(0.017)	(0.015)	(0.038)	(0.012)
Cons (CV at t=$t_{\tau,0}$)	4.928***	8.136***	22.042***	8.269***
	(0.400)	(0.351)	(0.890)	(0.271)
R-squared	0.817	0.763	0.919	0.825
N	4,970	8,347	4,580	8,347
Bigrams	497	835	458	835
Bigram FE	YES	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title
Mean	4.52	6.83	13.78	4.69
Rate of decline per year	**-0.004**	**-0.007**	**-0.016**	**-0.018**
	(0.003)	**(0.002)**	**(0.001)**	**(0.001)**
Years to zero CV	273.38	145.29	62.21	54.07
Panel B: Pioneer advantage				
	$Normalized\ Share_{n,\tau,t}$			
	(1) Industries	(2) Firms	(3) CBSAs	
$Pioneer_{n,\tau}$	6.504***	20.935***	1.321***	
	(1.751)	(4.883)	(0.254)	
$Pioneer_{n,\tau} * Years\ since\ emg_{\tau,t}$	-0.082	-0.489***	-0.035***	
	(0.074)	(0.185)	(0.011)	
R-squared	0.043	0.009	0.030	
N	1,515,850	49,854,895	7,751,024	
Bigrams	497	458	835	
Bigram FE	YES	YES	YES	
CBSA FE	YES	YES	YES	
Year FE	YES	YES	YES	
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	
Rate of decline per year	**-0.013**	**-0.023**	**-0.027**	
	(0.008)	**(0.004)**	**(0.003)**	
Implied years to zero advantage	79.32	42.81	37.74	

Notes: This table reports results from broadening regressions (in Panel A) and pioneer advantage regressions (in Panel B) along four dimensions: 1) industries, 2) occupations, 3) firms, and 4) locations (CBSAs). In Panel A, we regress the coefficient of variation calculated over $Normalized\ Share_{n,\tau,t}$ for each bigram and year where n is an industry (in column 1), occupation (in column 2), firm (in column 3), and location (in column 4). Panel B reports results from regressions of the $Normalized\ share_{n,\tau,t}$ on the pioneer status of n and the interaction of the pioneer status with the year since the technology bigram's emergence. As in Panel A, n is an industry (in column 1), firm (in column 2), and location (in column 3). The regressions use all technology bigrams that appear in at least 1000 job postings in our sample. All specifications are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. The normalized share is capped at 99th percentile of non-zero observations. All specifications indicate fixed effects used. Standard errors are clustered by Wikipedia title (technology). Note that the number of bigrams changes across specifications, depending on data availability on firms and industries in job postings. To test whether estimated coefficients are different across dimensions, we estimate stacked regressions using the same specifications as in Panels A and B, where we interact fixed effects with indicators for each dimension. In Panel A, the absolute rate of decline across CBSAs is 0.016 (0.004)***, 0.011 (0.003)***, and 0.003 (0.002) higher than across industries, occupations, and firms, respectively. In Panel B, the absolute rate of decline in pioneer advantage across CBSAs is 0.014 (0.007)* higher than across industries and 0.003 (0.005) higher than across firms. Similarly, the coefficient of $Pioneer_{n,\tau}$ is 19.614 (6.604)*** and 5.183 (1.659)*** higher for firms and industries than for CBSAs. For the coefficient on $Pioneer_{i,\tau} * Years\ since\ emg_{\tau,t}$, the estimated differences are 0.454 (0.247)* and 0.047 (0.702) between CBSAs and, respectively, firms and industries.

Table 10 – Robustness checks: Alternative samples and specifications

	Share of Top-5 CBSAs	Coefficient of Variation	log(Coefficient of Variation)	Normalized Share	Share College Educated
	(1)	(2)	(3)	(4)	(5)
	Concentration of Innovation	Region Broadening	Region Broadening by Skill	Rate of decline in Pioneer Persistence	Skill Broadening
	[*Table 3, col. 2*]	[*Table 4, Panel A, col. 3*]	[*Table 6, Panel A, col. 1, 2*]	[*Table 5, col. 3*]	[*Table 7, Panel A, col. 4*]
Estimate/Coefficient:	Share of Top-5 CBSAs	β_{RB}	$\beta_{RB}^{High\ skill} - \beta_{RB}^{Low\ skill}$	β_D/β_P	β_{SB}
Panel A: Influential patents					
Baseline: At least 1,000 cite-wt. patents	42.1%	-0.153***	-0.011***	-0.027***	-0.493***
		(0.012)	(0.003)	(0.003)	(0.036)
At least 1,250 cite-wt. patents	42.2%	-0.150***	-0.010***	-0.027***	-0.488***
		(0.012)	(0.003)	(0.003)	(0.037)
At least 1,500 cite-wt. patents	42.4%	-0.146***	-0.011***	-0.027***	-0.500***
		(0.012)	(0.003)	(0.003)	(0.038)
At least 1,750 cite-wt. patents	42.5%	-0.146***	-0.009***	-0.027***	-0.501***
		(0.013)	(0.003)	(0.004)	(0.040)
At least 2,000 cite-wt. patents	42.5%	-0.147***	-0.009***	-0.027***	-0.494***
		(0.013)	(0.003)	(0.004)	(0.040)
Panel B: Phrase Length					
Baseline: Bigrams	42.1%	-0.153***	-0.011***	-0.027***	-0.493***
		(0.012)	(0.003)	(0.003)	(0.036)
Bigrams and trigrams	42.0%	-0.157***	-0.011***	-0.025***	-0.494***
		(0.011)	(0.003)	(0.004)	(0.035)
Bigrams, trigrams, and unigrams (economically impactful only)	37.0%	-0.124***	-0.012***	-0.029***	-0.486***
		(0.008)	(0.003)	(0.003)	(0.033)
Panel C: Human Audit					
Baseline: Bigrams	42.1%	-0.153***	-0.011***	-0.027***	-0.493***
		(0.012)	(0.003)	(0.003)	(0.036)
Human-audited bigrams (economically impactful only)	44.4%	-0.142***	-0.014***	-0.025***	-0.574***
		(0.019)	(0.005)	(0.009)	(0.061)
Panel D: Alternative emergence years					
Baseline: At least 100 cite-wt. patents	42.1%	-0.153***	-0.011***	-0.027***	-0.493***
		(0.012)	(0.003)	(0.003)	(0.036)
At least 1 cite-wt. patent	41.9%	-0.157***	-0.013***	-0.026***	-0.465***
		(0.011)	(0.003)	(0.003)	(0.034)
At least 200 cite-wt. patents	42.2%	-0.154***	-0.012***	-0.028***	-0.511***
		(0.013)	(0.003)	(0.003)	(0.042)
50% of total cite-wt. patents	39.2%	-0.144***	-0.011***	-0.028***	-0.466***
		(0.009)	(0.003)	(0.008)	(0.027)
Panel E: Alternative weighting schemes					
Baseline: min(100, sqrt(# postings))	NA	-0.153***	-0.011***	-0.027***	-0.493***
		(0.012)	(0.003)	(0.003)	(0.036)
Unweighted regression	NA	-0.194***	-0.010***	-0.019***	-0.532***
		(0.014)	(0.002)	(0.005)	(0.043)

All bigrams with job postings	NA	-0.209***	-0.009***	-0.020***	-0.490***
		(0.014)	(0.002)	(0.003)	(0.046)
Log-wt: min(100, log(# postings))	NA	-0.178***	-0.011***	-0.022***	-0.515***
		(0.013)	(0.003)	(0.004)	(0.039)
Bigrams collapsed into technologies	NA	-0.156***	-0.013***	-0.026***	-0.490***
		(0.010)	(0.002)	(0.005)	(0.033)
100+ normalized EC counts	NA	-0.149***	-0.013***	-0.024***	-0.532***
		(0.015)	(0.004)	(0.006)	(0.044)

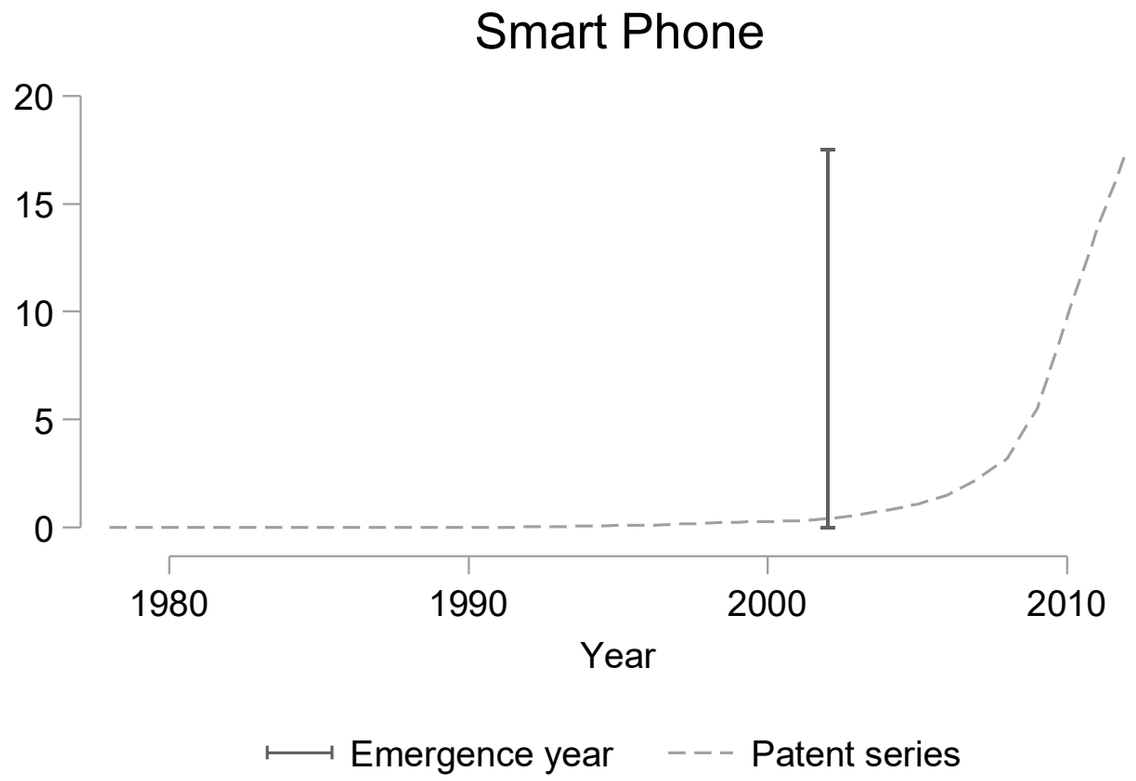
Notes: This table reports robustness checks to our primary ("Baseline") results. Panel A reports robustness to changing the threshold for defining bigrams associated with "influential innovations." In the baseline, we retain only those that appear in patents accumulating a total of at least 1,000 weighted citations. The panel shows four variations, with cutoffs ranging from 1,250 to 2,000 citations. Panel B presents results from extending our sample to include technology trigrams and unigrams. In the baseline, we include only technology bigrams. The row "Bigrams and trigrams" includes trigrams along with technology bigrams in the analysis, while the row "Bigrams, trigrams, and unigrams (economically impactful only)" further adds unigrams. In Panel C, "Human audited bigrams (economically impactful only)", we rely on human reading to determine whether or not a bigram describes a technology, instead of the Wikipedia filter, reporting results from including only those technology bigrams that survived the human auditing process. Panel D reports results from variations in defining emergence years. In the baseline, the emergence year is defined as the first year in which (a) 100 citation-weighted patents associated with that technology had been already applied for and (b) where the next five years had 10% annual growth in (smoothed) weighted patenting. The rows "At least 1 cite-wt. patent" and "At least 200 cite-wt. patents" explore changing the threshold from 100 cite-weighted patents in (a) to at least one cite-weighted patent and (b) at least 200 cite-weighted patents, respectively. The row "50% of total cite-wt. patents" changes our definition of emergence years completely: the emergence year of a given bigram is defined as the first year when 50% of maximum peak of citation-weighted patent counts is realized. Panel E presents robustness to changing weighting schemes in regressions. Baseline regressions are weighted by the square root of the total number of job postings mentioning that technology in that year, capped at 100. The row "Unweighted regression" replicates baseline regressions with equal weights for each observation. The row "All bigrams with job postings" performs unweighted regressions with all bigrams that are mentioned by at least one job posting. The row "Log-wt" weights observations by the log of the number of postings observed for each technology bigram in a given year, capped at 100. The row "Bigrams collapsed into technologies" replicates our results when all bigrams associated with a given Wikipedia title (technology) are collapsed into the technology. The last row, "100+ normalized EC counts" replicates our results with bigrams that cumulate more than 100 normalized earnings calls mentions. The rows reporting unigrams in Panel B and human-audited bigrams in Panel Creport results only using economically impactful technologies. See the original regressions for full details.

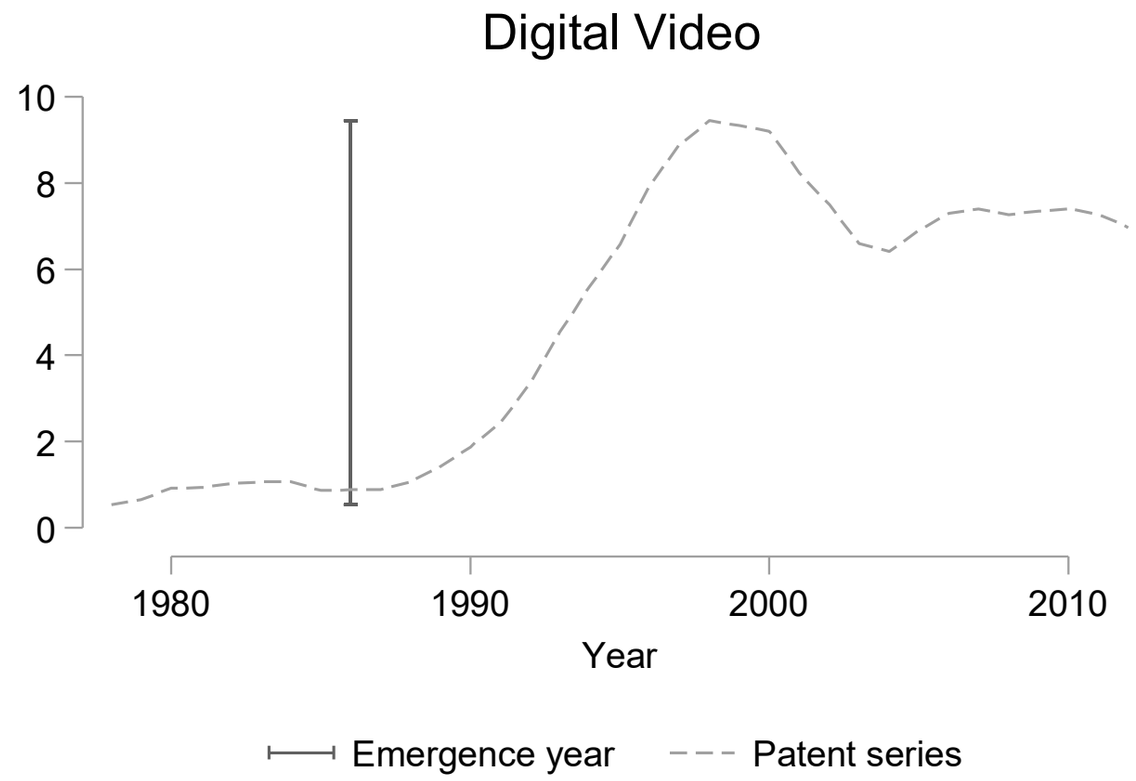
Table 11 – Robustness checks: Alternative specifications of standard errors

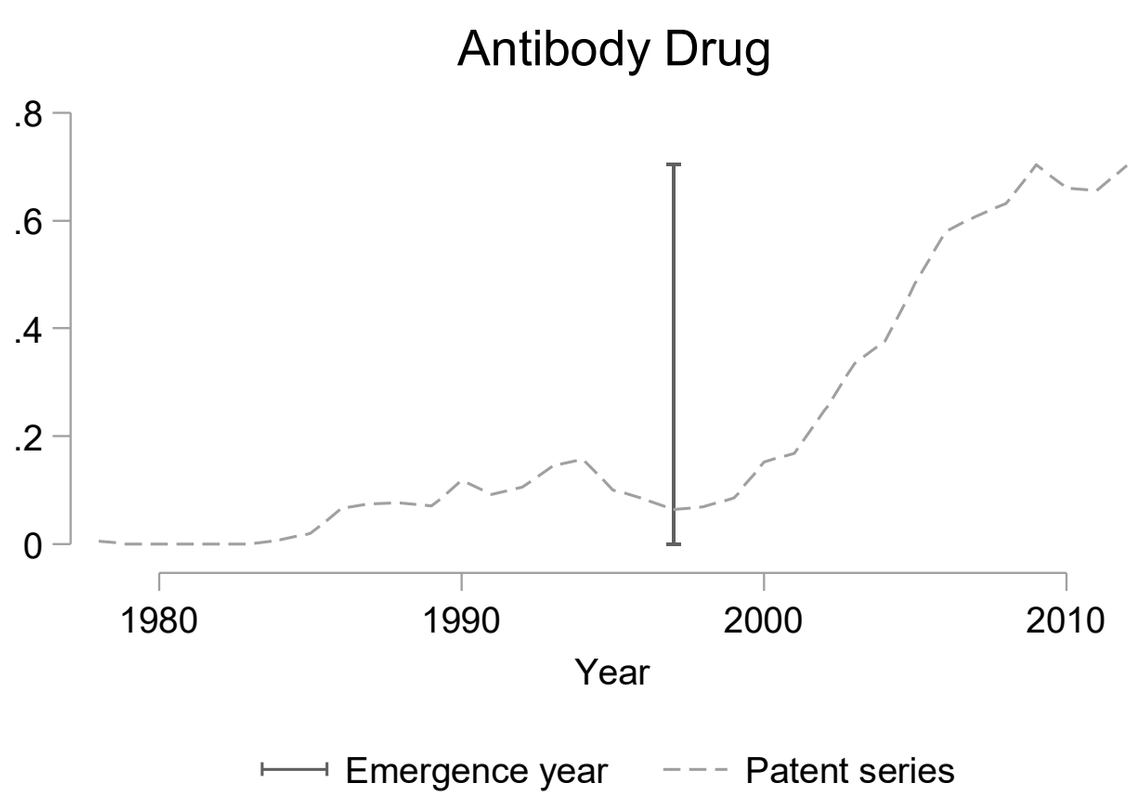
	Coefficient of Variation	Coefficient of Variation	Normalized Share	Share College Educated
	(1)	(2)	(3)	(4)
	Region Broadening	Region Broadening by Skill	Pioneer Persistence	Skill Broadening
	[*Table 4, Panel A, col. 3*]	[*Table 6, Panel A, col 1,2*]	[*Table 5, col. 3*]	[*Table 7, Panel A, col. 4*]
$Years\ since\ emergence_{\tau,t}$ (*High Skill* in col. 2)	-0.153	-0.027		-0.493
[*baseline*] Cluster, Wikipedia Title level	(0.012) ***	(0.002) ***		(0.036) ***
Cluster, Bigram level	(0.009) ***	(0.002) ***		(0.028) ***
Cluster, Year level	(0.018) ***	(0.004) ***		(0.042) ***
Bootstrap (500 replications)	(0.009) ***	(0.002) ***		(0.027) ***
$Years\ since\ emergence_{\tau,t}(Low\ Skill)$		-0.038		
[*baseline*] Cluster, Wikipedia Title level		(0.003) ***		
Cluster, Bigram level		(0.002) ***		
Cluster, Year level		(0.005) ***		
Bootstrap (500 replications)		(0.002) ***		
$Pioneer_{i,\tau}$			1.321	
[*baseline*] Cluster, Wikipedia Title level			(0.254) ***	
Cluster, Bigram level			(0.213) ***	
Cluster, Year level			(0.059) ***	
Cluster, CBSA level			(0.203) ***	
Cluster, State level			(0.190) ***	
Cluster, CBSA-Wikipedia Title levels			(0.277) ***	
Bootstrap (500 replications)			(0.214) ***	
$Pioneer_{i,\tau} * Years\ since\ emergence_{\tau,t}$			-0.035	
[*baseline*] Cluster, Wikipedia Title level			(0.011) ***	
Cluster, Bigram level			(0.009) ***	
Cluster, Year level			(0.003) ***	
Cluster, CBSA level			(0.007) ***	
Cluster, State level			(0.006) ***	
Cluster, CBSA-Wikipedia Title levels			(0.011) ***	
Bootstrap (500 replications)			(0.008) ***	
$\beta(Pioneer_{i,\tau} * Years\ since\ emergence_{\tau,t})$ $/\beta(Pioneer_{i,\tau})$			-0.027	
[*baseline*] Cluster, Wikipedia Title level			(0.003) ***	
Cluster, Bigram level			(0.003) ***	
Cluster, Year level			(0.001) ***	
Cluster, CBSA level			(0.003) ***	
Cluster, State level			(0.003) ***	
Cluster, CBSA-Wikipedia Title levels			(0.004) ***	
Bootstrap (500 replications)			(0.003) ***	
Bigram FE	YES	YES	YES	YES
Skill FE	NA	YES	NA	NA
CBSA FE	NA	NA	YES	NA
Year FE	NA	NA	YES	NA

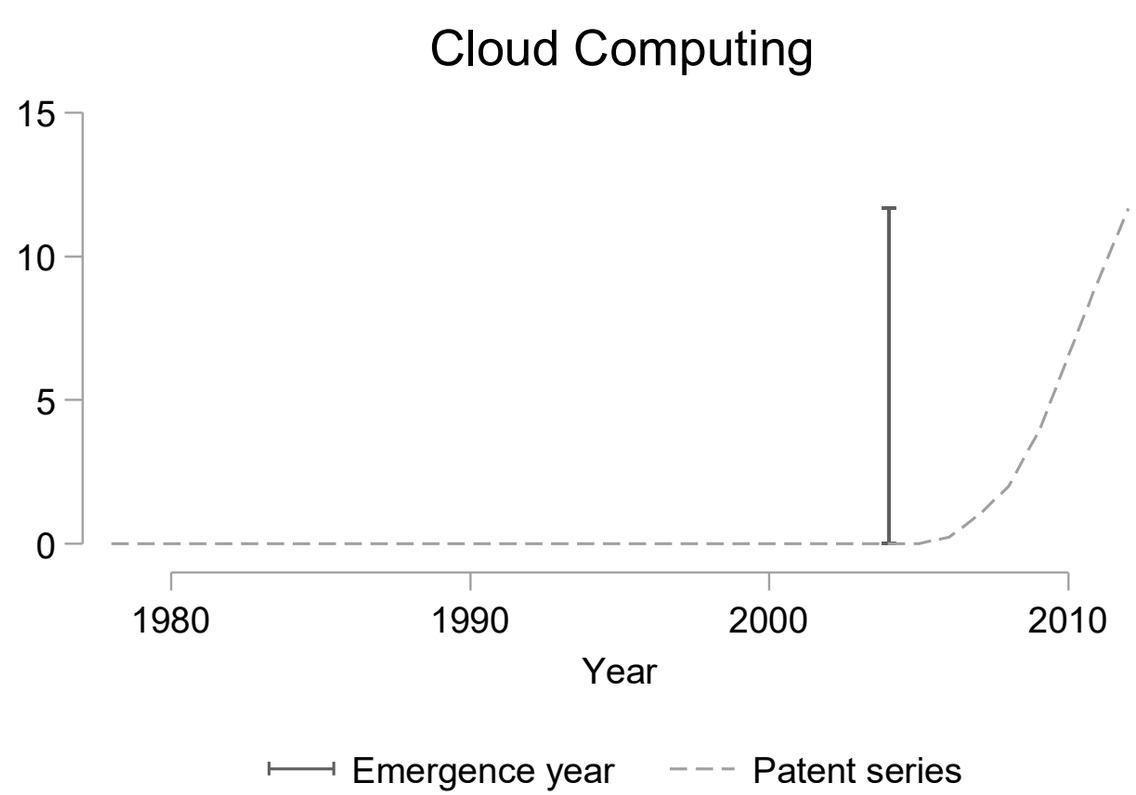
Notes: This table the reports results from varying specifications for standard errors corresponding to coefficient estimates for our main results – region broadening, region broadening by skill, pioneer persistence and skill broadening. The statistical significance of coefficients is indicated by the asterisks next to each parenthesis. For the results in Columns 1, 2, and 4, we report standard errors clustered at the Wikipedia title level (baseline), bigram level, and year level. In Column 3, we report standard errors clustered at the Wikipedia title level (baseline), bigram level, year level, CBSA level, state level, and CBSA x Wikipedia title level (double-cluster). To cluster CBSAs into the state level, we assign CBSAs that are shared by more than one state to the state with lowest FIPS number. For each result, in the last row, we report bootstrapped standard errors for each specification. Bootstrapped standard errors are computed based on 500 replications with replacement from the original sample. Re-sampling was done at the bigram-level (sampling bigram-blocks with ten years of observations). See the original regressions for full details.

Figure 1– Examples of emergence year definition



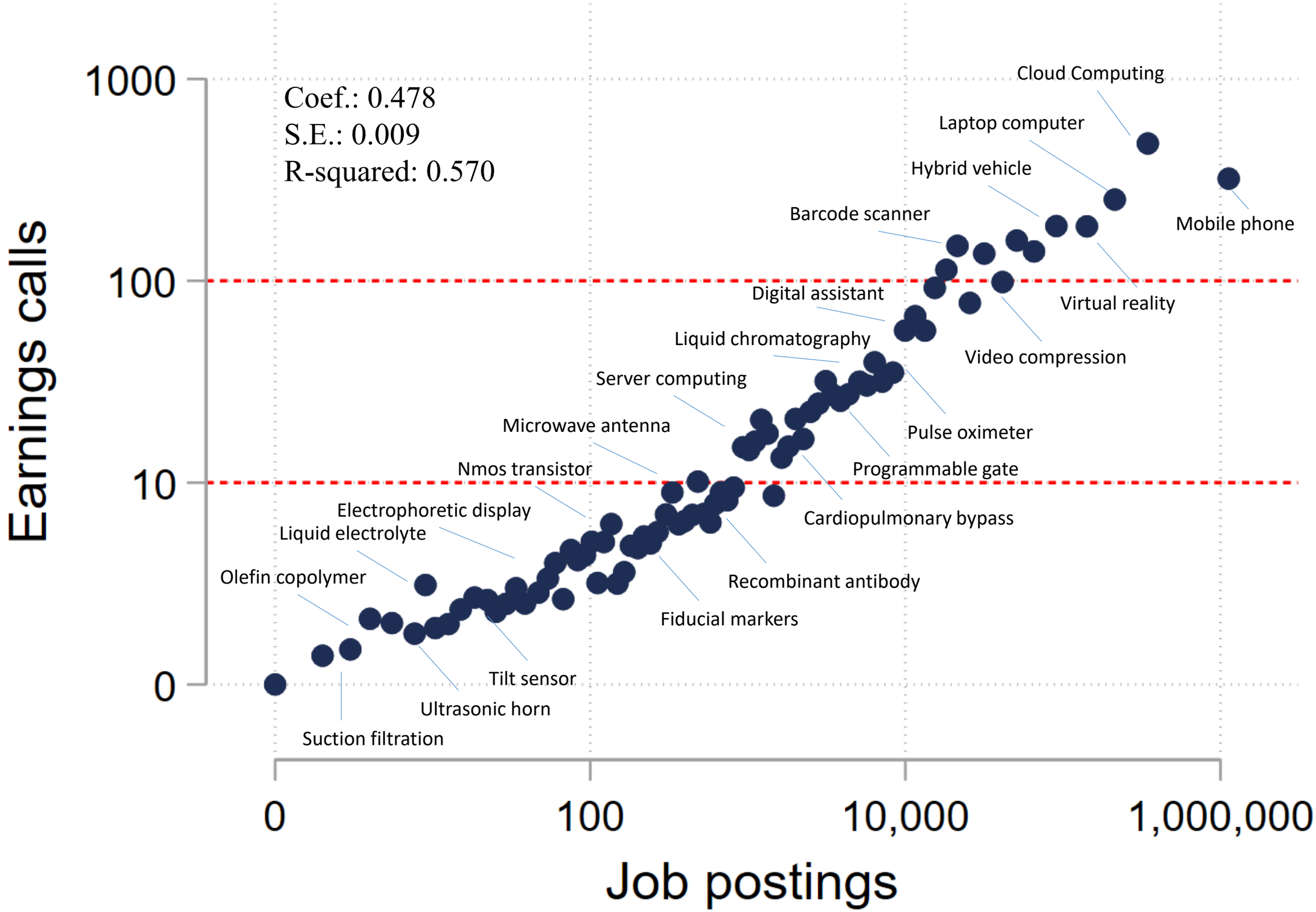






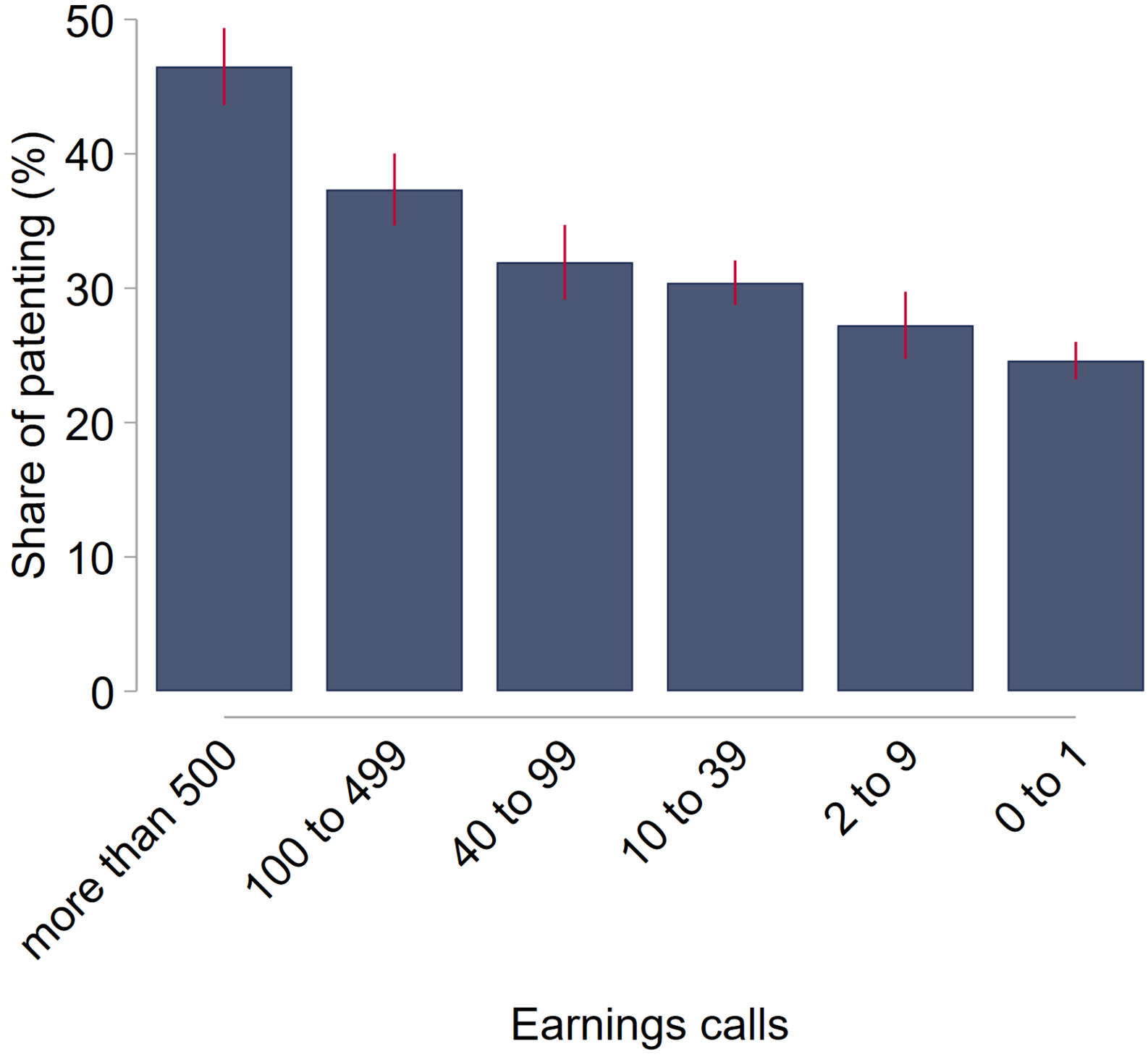
Notes: The figure shows four examples of the attribution of emergence years. In each example, the time series plots the smoothed number of cite-weighted patents associated with the technology by year of application of the patent. For each bigram, we mark the emergence year as the first year in which (a) the technology reaches 100 cite-weighted patent applications and (b) where the next five years had at least 10% annual growth in the (smoothed) series for each bigram. For more details, refer to Section 2.c.

Figure 2 – Earnings calls and job postings for new technologies



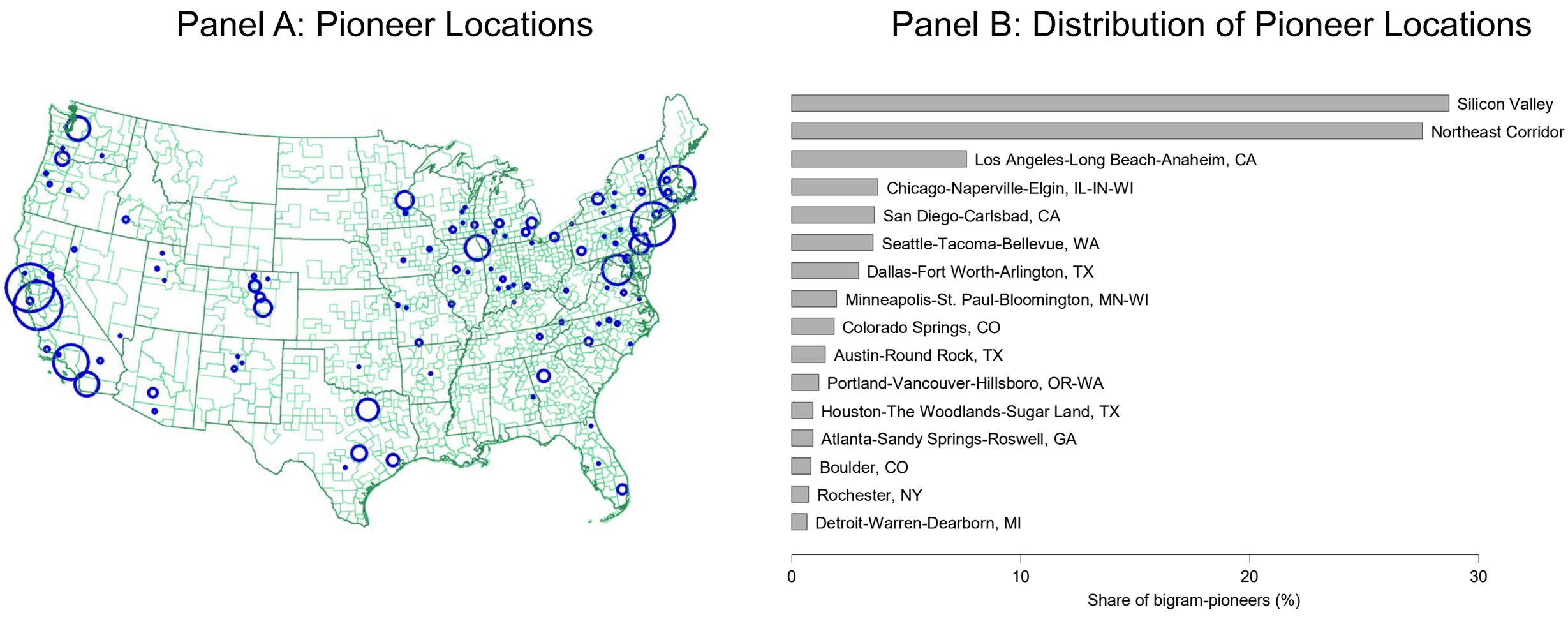
Notes: The figure shows a binned scatterplot at the technology bigram level of the number of the number of earnings calls that mention a given technology (y-axis) against the number of job postings that mention the technology bigram (x-axis). Some examples are labeled next to their bins.

Figure 3-- Share of patents mentioning a new technology filed in Silicon Valley, New York, Seattle and Boston, by technology's economic importance (earnings calls mentions)



Notes: The figure shows results from a regression of concentration at the bigram-level on indicators for the number of earnings calls that mention a certain bigram, with the respective 95% confidence interval in red. For each bigram, concentration is measured by the share of patenting associated with that bigram in the top five CBSAs (San Jose-Sunnyvale-Santa Clara, CA; San Francisco-Oakland-Hayward, CA; New York-Newark-Jersey City, NY-NJ-PA; Seattle-Tacoma-Bellevue, WA; Boston-Cambridge-Newton, MA-NH). The top five CBSAs are the five regions with the highest number of patents associated with technologies with more than 100 earnings call mentions. Standard errors are clustered by Wikipedia title (technology). Using a F-test, we reject the hypothesis that all coefficients are equal, with a p-value of 0.000 (F-statistic of 42.72).

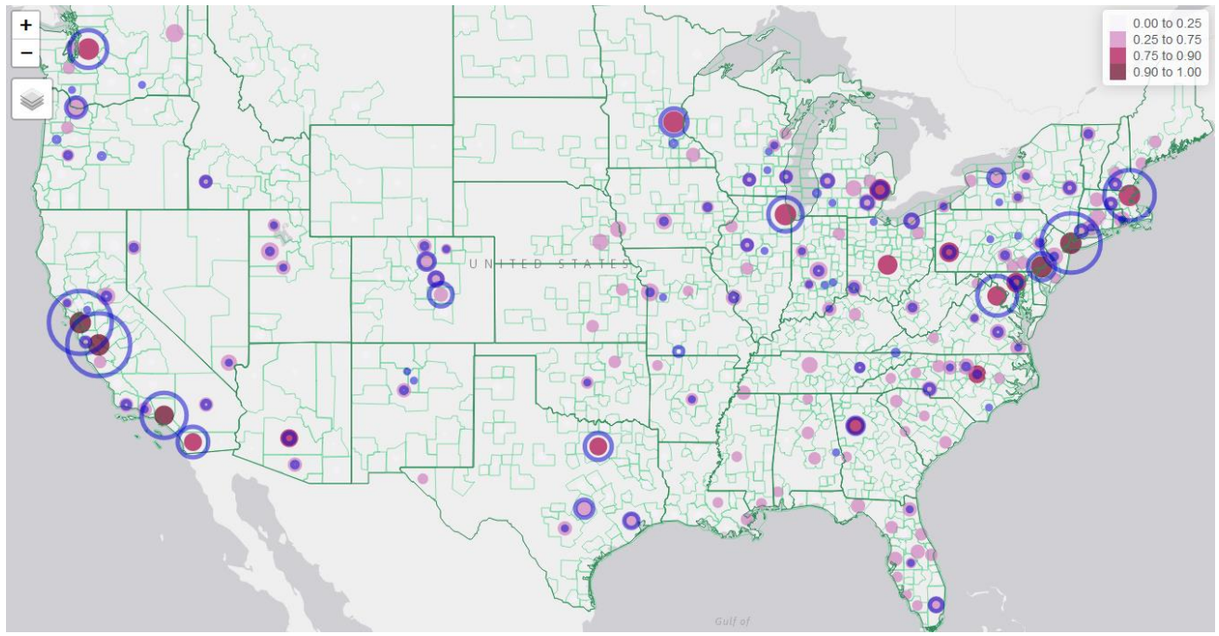
Figure 4 – Distribution of pioneer locations



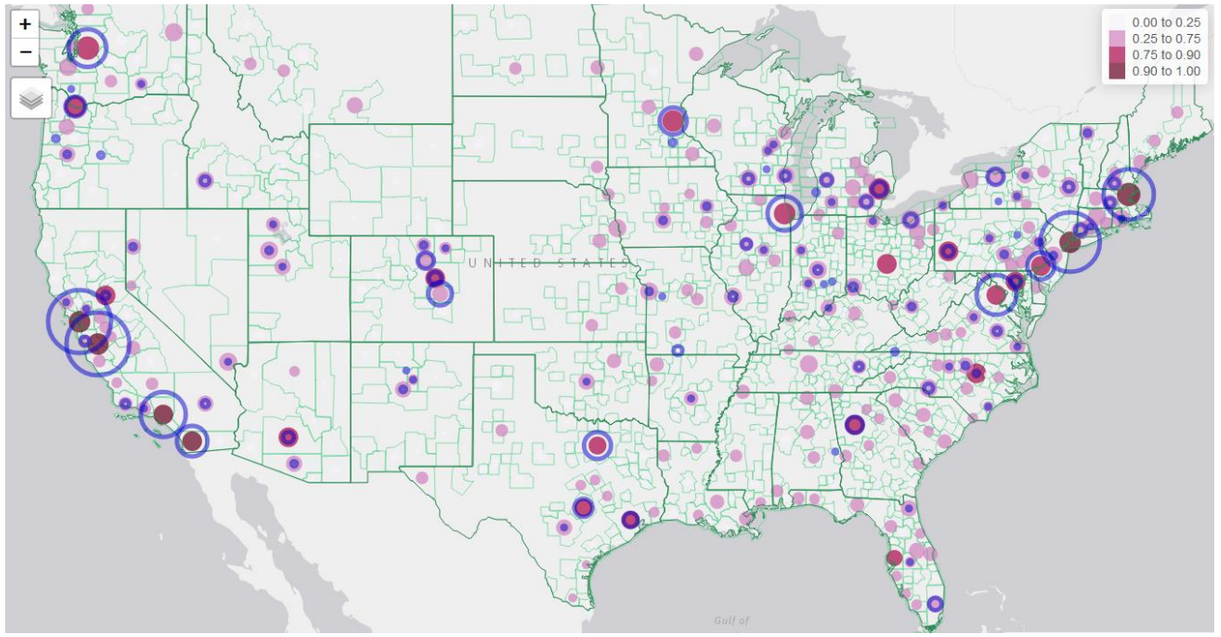
Notes: This figure shows the distribution of pioneer CBSAs. Panel A displays as blue circles CBSAs that are pioneer locations for at least one bigram with more than 100 earnings call mentions. The size of the circles is proportional to the share of technology bigrams for which the CBSA is a pioneer location. Panel B shows a plot of the percentage of technology bigram-pioneer location pairs accounted for by each CBSA, for the top 20 CBSAs. We combine the CBSAs San Jose-Sunnyvale-Santa Clara, CA and San Francisco-Oakland-Hayward, CA, and label the region as Silicon Valley. Similarly, we combine New York-Newark-Jersey City, Boston-Cambridge-Newton, Washington-Arlington-Alexandria, and Philadelphia-Camden-Wilmington, and label the region as the Northeast Corridor.

Figure 5 – Geographic diffusion of technology job postings, by year since emergence

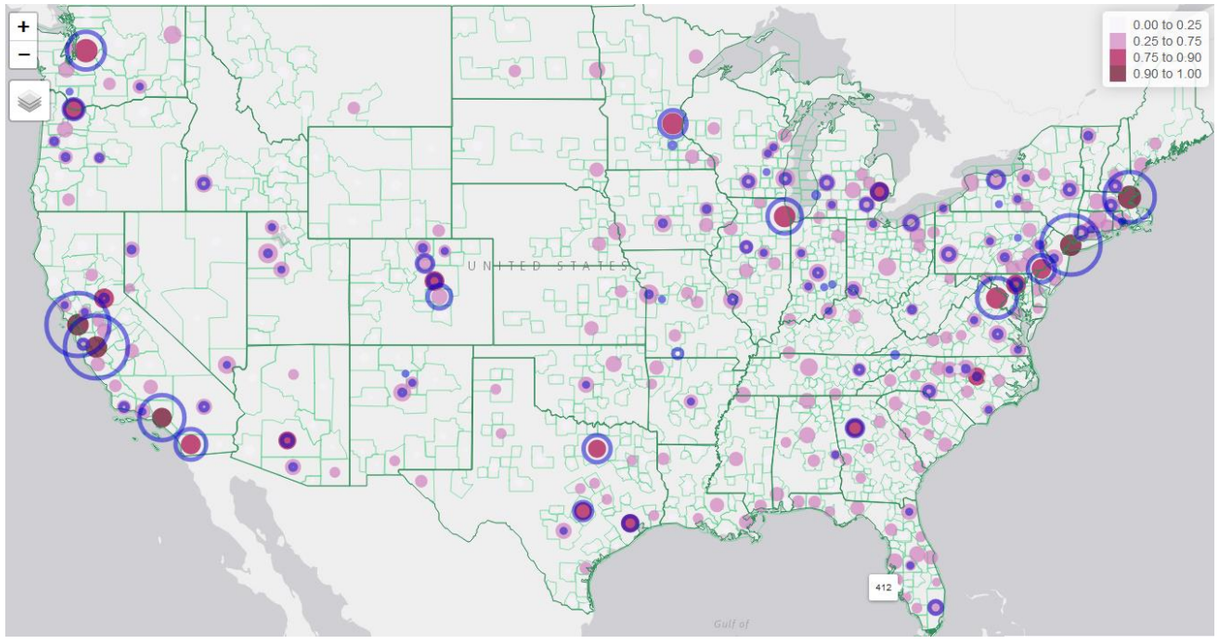
Years since emergence: 0-5



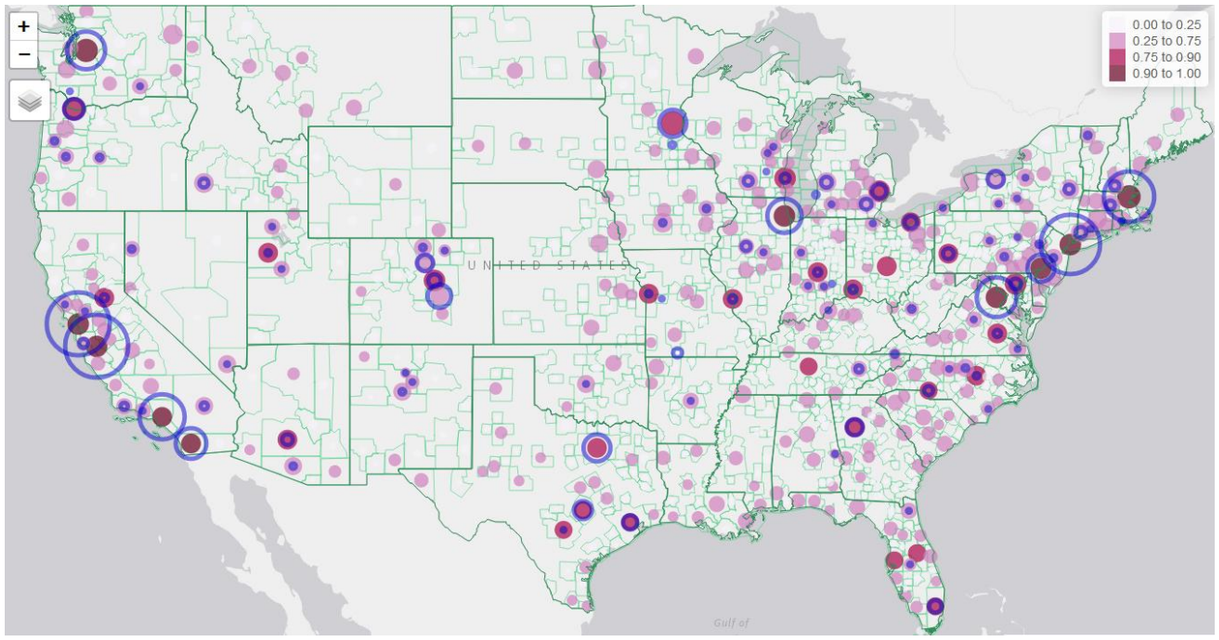
Years since emergence : 6-10



Years since emergence: 11-20

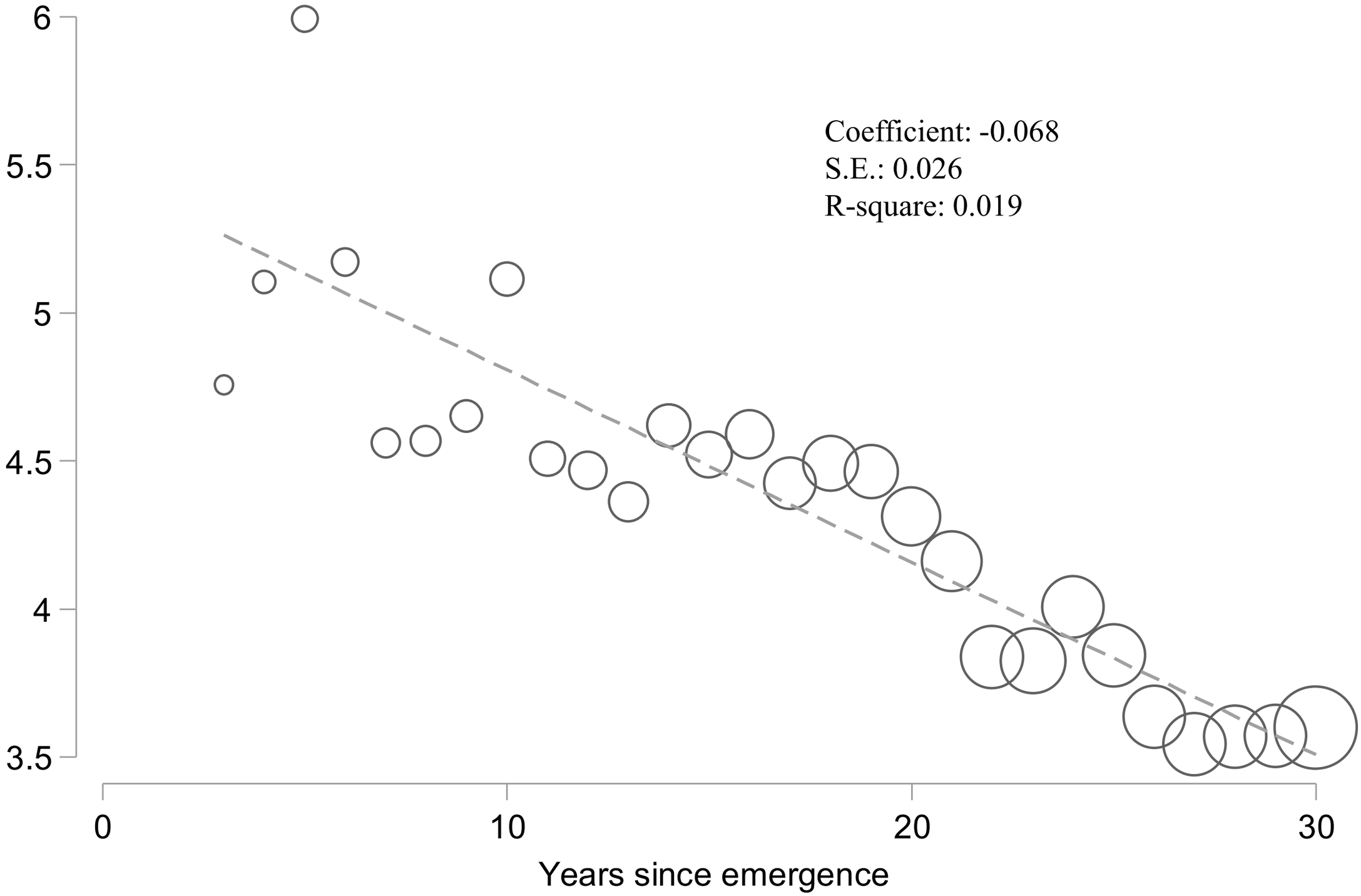


Years since emergence : 21-30



Notes: This figure plots maps with pioneer locations and job postings associated with technology bigrams by year since technology bigram emergence. Pioneer locations are marked with solid blue circles and technology job postings are in solid purple circles. For each CBSA and emergence year, we calculate the share of technology bigrams for which the CBSA records a non-negligible presence of technology jobs ($Normalized\ share_{c,\tau,t} \geq 10\%$) and denote a higher share of technology bigrams with a darker color. For example, the first map plots the share of technologies with a normalized share of technology job postings greater than 10% for each CBSA between zero and five years since the emergence of the technology. The second map replicates this picture for six to ten years after emergence, and so forth. The sample for this map only contains technologies that appear in at least 100 earnings calls.

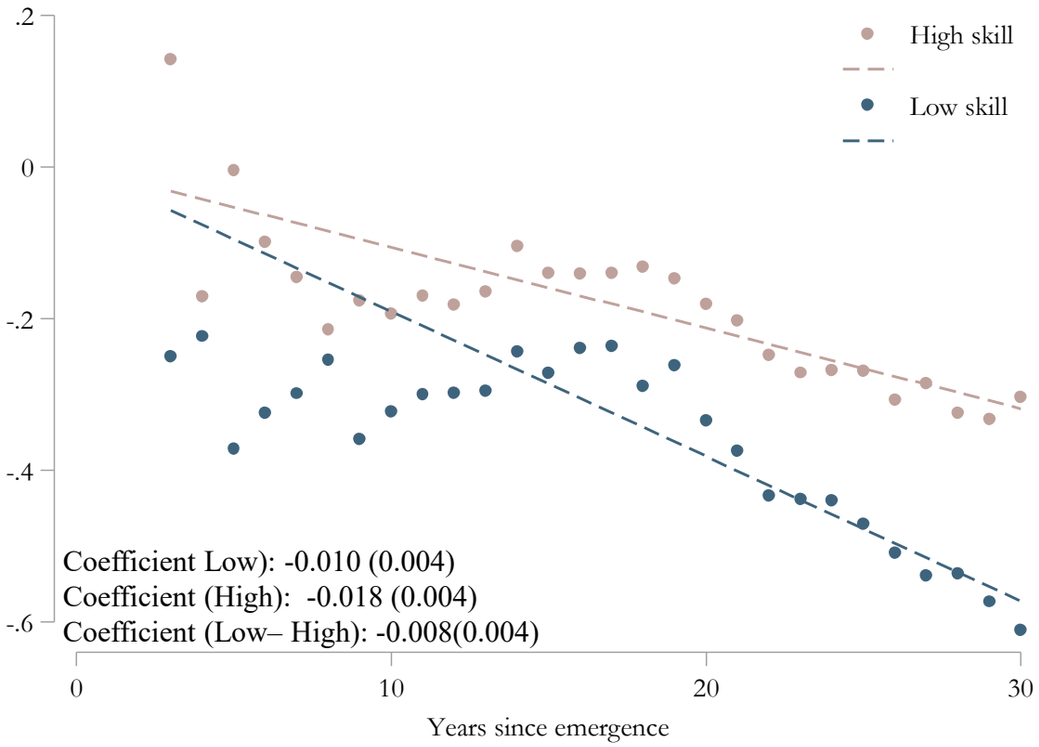
Figure 6 – Geographic concentration of technology job postings across CBSAs, by year since emergence



Notes: This figure shows a binned scatter plot at the technology bigram x year level of the coefficient of variation (CV) of the normalized share of technology job postings over time. We calculate the CV of the normalized share of technology job postings by dividing the standard deviation of *Normalized Share*$_{c,\tau,t}$ across locations c in year t by its mean in year t for each technology bigram τ. Each dot represents the weighted average of the CV (calculated across technologies) for each year since emergence, where the weight is the square root of the number of job postings for a bigram in a year, capped at 100. The circle sizes are proportional to the same weight. The regression line in the plot corresponds to a regression of the CV on year since emergence, as in Table 4, Panel A, column 1. We only include technology bigrams that appear in at least 100 earnings calls. Observations in and after the year of emergence are included.

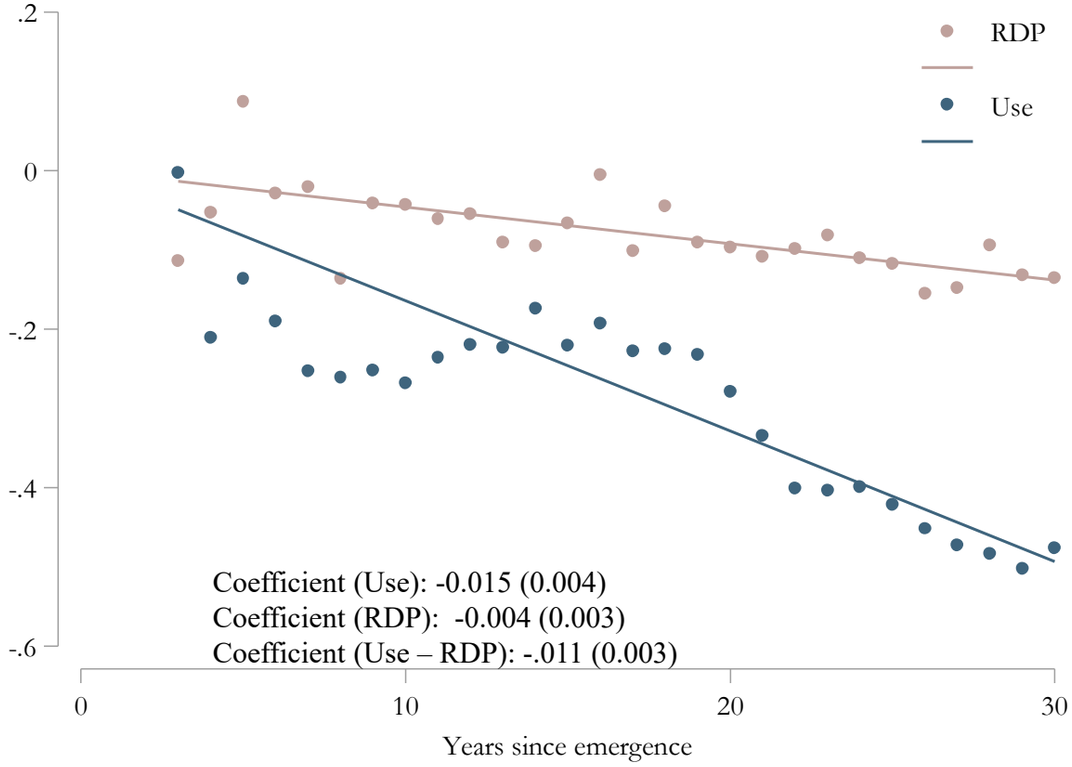
Figure 7 – Geographic concentration relative to year since emergence, by skill and type of job posting

Panel A: Geographic concentration by skill



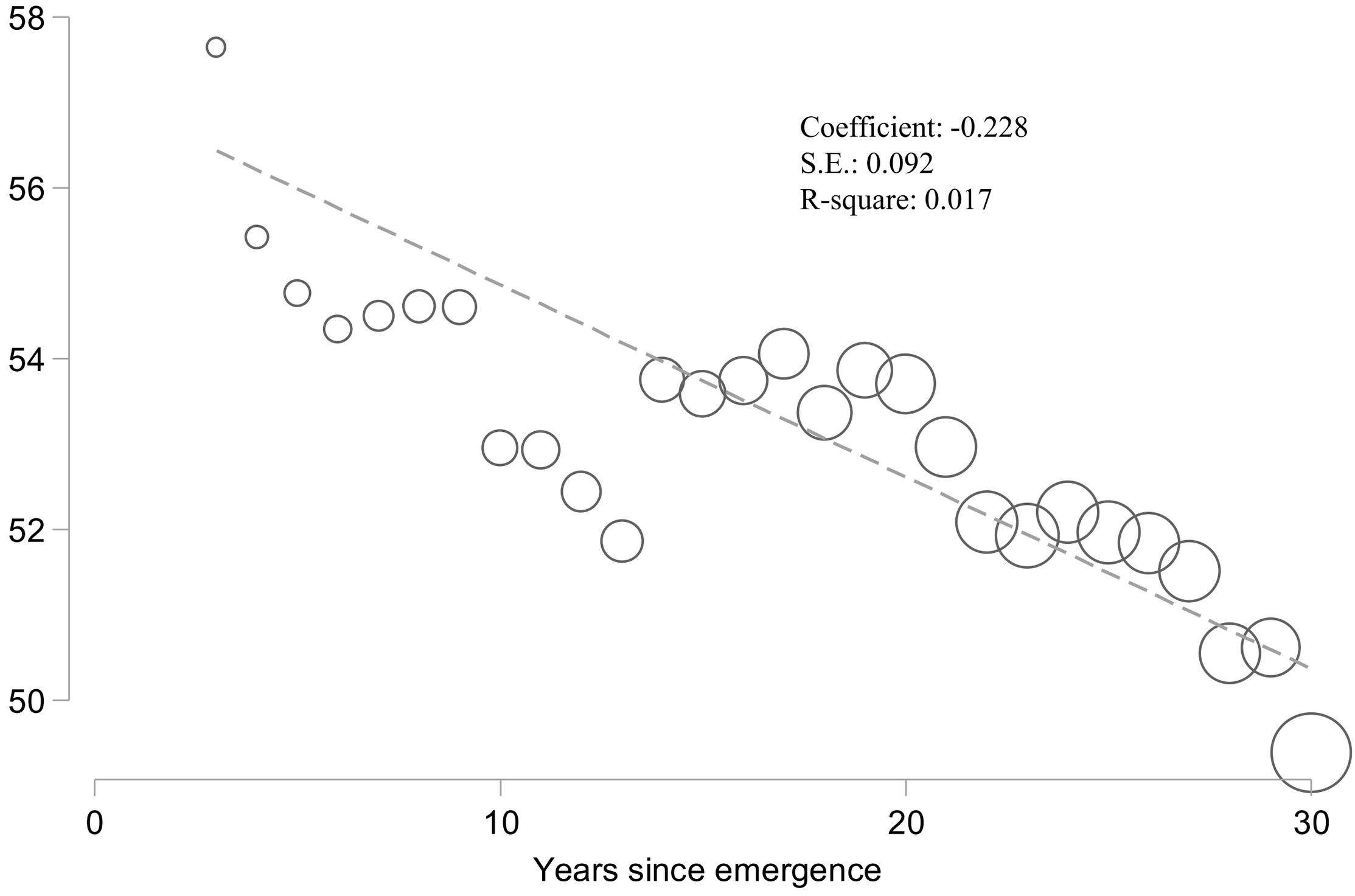
Notes: This figure plots a binned scatter plot at the technology bigram x year x skill category level of the log of the coefficient of variation (CV) of the normalized share of bigram job postings relative to the year since emergence, by skill-level of job posting (high and low). The CV is calculated as the ratio of standard deviation to the mean of the normalized share across CBSAs for each technology bigram x year x skill category triplet. The red dots represent high-skill postings, and the blue dots represent low-skill postings. The fitted lines weigh observations by the square root of the total number of postings for a technology bigram in a year, capped at 100.

Panel B: Geographic concentration by use/research, develop and production



Notes: This figure plots a binned scatter plot at the technology bigram x year x job type level of the log of the coefficient of variation (CV) of the normalized share of bigram job postings relative to the year since emergence, by job type (RDP and use). The CV is calculated as the ratio of standard deviation to the mean of the normalized share across CBSAs for each technology bigram x year x job type triplet. The red dots represent RDP-related postings, and the blue dots represent use-related postings. The fitted lines weigh observations by the square root of the total number of postings for a technology bigram in a year, capped at 100.

Figure 8 - Share of technology job postings requiring a college education, by year since emergence



Notes: This figure shows a binned scatter plot at the technology bigram x year level of the share of technology postings requiring a college education by year since emergence. The share of college-educated postings for each technology bigram x year observation is measured as discussed in Section 5. Each dot represents a weighted average over technology bigrams of the share for each year since emergence, where the weight is the square root of the number of postings for a bigram in a year, capped at 100. The circle sizes are proportional to the same weight. The regression line in the plot corresponds to a regression of share of college-educated postings on the year since emergence as in Table 7, Panel A, column 1. We only include technology bigrams that appear in at least 100 earnings calls.

Internet Appendix for

The Diffusion of New Technologies

Aakash Kalyani, Nicholas Bloom, Marcela Carvahlo, Tarek Hassan,
Josh Lerner, and Ahmed Tahoun

June 20, 2024

Data Appendix

We process four sources of text data, and then combine them with U.S. government data to conduct our analyses.

1. Text Sources

1.1 Patents

We download two separate sets of data for about six million utility patents, applied for at the U.S. Patent and Trademark Office (USPTO) between 1976 and 2014 and granted by 2018. First, we download full patent text XML files from the USPTO website. Second, we download processed patent variables, such as assignee names, inventor names and location, application and award year, citations (through 2018), and the primary CPC class, all from PatentsView.org. We restrict our sample to the sample of patents filed by U.S. inventors, about one-half of these awards.

We map the FIPS county identifier provided for inventors of each patent to Core-Based Statistical Areas (CBSAs), using a crosswalk provided by the U.S. Census Bureau. For patents with multiple inventor CBSAs, we assign the patent to each CBSA. We also standardize citation counts to control from truncation and time differences: we divide citations for each patent by the average number of citations for patents with a primary assignment to the same four-digit Combined Patent Classification (CPC) patent class and application year.

A typical patent award has six text sections: (1) title, (2) abstract, (3) background, (4) summary, (5) detailed description, and (6) claims. After we remove stop words (such as "of," "the," and "from") following Kelly et. al. (2021) and Gentzkow et. al. (2019), we combine text from all of these sections into one large text string by appending all available text sections. We then break this large text string down into two-word combinations or bigrams. During this process, we use only those bigrams which are mentioned at least twice in a patent.

1.2 Corpus of Historical American English

The Corpus of Historical American English (COHA) is a collection of 116,759 documents published between 1880 and 1970. We download COHA from www.english-corpora.org/coha. These

include fiction and non-fiction books and newspaper and magazine articles. As with patents and earnings calls, we decompose these documents into about 400 million unique bigrams.

1.3 Wikipedia pages

For each of our 36,563 novel and influential bigrams, we search on Wikipedia using the Wikipedia search functionality. We then download the full text of the first suggested Wikipedia page using the Wikpedia Python API (Wikipediaapi). For each of these Wikpedia pages, we store separately the title, the section headings, and the text of each section. While counting bigrams in Wikipedia, we count singular, plural, and unigram versions of these bigrams. For example, when counting `smart phone', we include counts for "smart phone," "smarts phone," "smart phones," "smartphone," "smartphones," "smartsphone," and "smartsphones." We follow the same approach when we count these bigrams in other corpora.

1.4 Earnings conference call transcripts

From Refinitiv EIKON, we collect the complete set of 321,189 English-language transcripts of earnings conference calls held from 2002 through 2019. Out of these, we drop 5,552 transcripts because we could not reliably match them to a company name in Compustat. We obtain a total of 11,992 firms and 301,294 firm x quarter observations. We count our bigrams in the full text of these earnings calls and collect the number of earnings calls that each of our bigrams is mentioned in.

1.5 Burning Glass job postings

From Burning Glass (BG), we obtain about 200 million job postings posted online in the U.S. Similar to patents, job postings data comes in two sets. The first contains the full text coded in XML files, while the second contains processed information about each job posting (such as occupation codes). We undertake minimal processing of job postings' textual data: (1) removing non-letter sections of job postings; (2) removing the top 50 and bottom 50 words from each job posting, as mentioned in Section 2; and (3) as a consequence of step (2), excluding any job posting with less than 100 words. We then perform word counts over the remaining text. As before, while counting bigrams in Burning Glass, we count singular, plural and unigram versions of these bigrams. For example, when counting "smart phone," we include counts for "smart phone,"

"smarts phone," "smart phones," "smartphone," "smartphones," "smartsphone," and "smartsphones." We follow the same approach when we count these bigrams in other corpora.

2. Merging Burning Glass Occupations with American Community Survey (2015)

We obtain occupation and location demographic variables from the 2015 American Community Survey (ACS), downloaded on March 9, 2020. We examine respondents who are at least 25 years old, and report at least one year of schooling and a non-zero annual wage. We calculate the "share of college-educated people" in a particular occupation by dividing the number of people who report a particular occupation and have at least three years of college education by the total number of people who report the occupation. We calculate the "share of post-graduates" in a particular occupation by dividing the number of people who report a particular occupation and have at least a masters' degree by the total number of people who report the occupation. We calculate the average wage in the occupation by taking an average over all annual incomes of people reporting a particular occupation. As for locations, we calculate skill levels using reported locations in the ACS and following the same methodology as for occupations. We also obtain population data for each CBSA from the ACS by performing a sample-weighted count of people who reported to live in a certain CBSA.

We merge the occupation level data from the ACS with occupation level aggregates in BG using six-digit SOC codes. Data on some six-digit SOC codes are reported in aggregated form in the ACS: for example, data on the occupational code 17-2021 (agricultural engineers) are reported as 17-20XX, along with class 17-2031 and others. In these cases, we map the six-digit SOC codes in BG to their aggregated values in the ACS.

As we do for occupations, we calculate share of college-educated people for CBSAs by dividing the number of people who report a particular CBSA as their residence in the ACS and have at least three years of college education by the total number of people in the CBSA. Other skill measures are calculated similarly.

3. Matching job postings to occupations, locations, industries and firms

Burning Glass codes job postings into occupations (Standard Occupational Classification (SOC) codes), locations (counties), and industries (North American Industrial Classification Codes). BG also extracts an employer name with these job postings.

There are 836 occupations with six-digit SOC codes and 312 industries with NAICS Codes in the sample. We map counties in the BG data to CBSAs and use them as the unit for our geographical analysis. We do so by using the crosswalk made available by National Bureau of Economic Research (NBER). In this process, we lose about 2.8% of the job postings.

In order to assign firms to job postings, we use the employer strings provided by BG, which are available for 42.1% of job postings. Furthermore, these employer strings are not standardized or cleaned. For example, there are employer strings of the form "Tesla Motors Gigafactory," "About Tesla," and "Tesla Incorporated." We generate firm identifiers from these raw employer strings using a modification of the process in Autor et. al. (2020):

1) We search the raw employer string on Bing.com and store the top five search result links. For instance, for the employer string "Tesla Incorporated," we get https.www.tesla.com, https://en.wikipedia.org.wiki/tesla-inc, https://www.britannica.com/topic/tesla-motors, https://www.bloomberg.com/quote/tsla/us, https://www.marketwatch.com/investing/stock/tsla.
2) We group two employer strings under a single identifier if they share at least two out of top five links in common with each other.

Using this process, we group together 477,583 employer strings in BG into 329,158 unique firm identifiers.

We also match these employers to patent assignees using string matching. We implement the following modified "Term Frequency — Inverse Document Frequency" (tf-idf) algorithm. To do so, we:

a. Decompose employer and assignee strings into 5 letter combinations. For example, "Alphabet" is broken into: "alpha," "lphab," "phabe," and "habet."
b. Calculate a term frequency, which is the frequency of the five letter combination in the string. In our example of "Alphabet," each combination uniquely appears in the strings.

We calculate an inverse document frequency (idf), which is inverse of the frequency with which the combination appears in all strings of assignees and BG employers.

c. We combine the term frequency (tf) with an inverse document frequency (idf) to obtain a vector of combinations for each string:

$$v_{s,c} = tf_{c,s} * idf_c$$

where $tf_{c,s}$ is the term frequency of the 5-letter combination c in string s, idf_c is the inverse document frequency of each combination, and $v_{c,s}$ is the value attributed to each combination separately for every string.

d. Finally, we normalize each vector $\boldsymbol{v}_s$ so that the norm is 1. We then calculate similarities between two strings s and s' using dot product of their respective normalized vectors.

$$d_{s,s'} = \boldsymbol{v}_s . \boldsymbol{v}_{s'}$$

e. We match two strings if $d_{s,s'} \geq 0.75$.

A human audit of these matches resulted in an 86% accuracy rate.

4. University Data

We download data on U.S. research universities from the U.S. National Science Foundation's Higher Education Expenditure on R&D (HERD) survey, which collects detailed statistics on research expenditure by these universities, and from the Integrated Postsecondary Education Data System (IPEDS) surveys provided by the U.S. Department of Education's National Center for Education Statistics (NCES). From these datasets, we construct the following variables:

1) Number of research universities in a CBSA: HERD provides details of universities which spend more than $150,000 in research. We map university zip codes, provided as a part of university addresses, to CBSAs using a crosswalk provided by U.S. Census Bureau. Finally, we count the number of research universities in a CBSA to construct our variable.

2) University assets in a CBSA: IPEDS provides details of finances for most post-secondary educational institutions in the U.S. As with 1) above, we assign these universities to CBSAs and then aggregate their assets over CBSAs.

Appendix Tables and Figures

Appendix Table 1 – Examples of novel influential bigrams from patents that pass and fail the Wikipedia technology bigram criteria

Technologies (pass)	Non-technologies (fail)
antigen binding	account manager
based solutions	active directory
branched chain	airway pressure
cell lines	aqueous phase
cloud computing	business model
communication channel	customer care
computer network	customer relationship
data quality	customer requests
data sources	dosage form
disk drive	email address
distributed computing	hardware software
fiber optic	healthcare provider
global positioning	homeland security
internet explorer	host computer
machine learning	identity theft
microsoft office	image data
mobile devices	management software
monoclonal antibody	mission critical
multiple access	performance management
optical fiber	performance metrics
personal computer	pharmaceutical composition
polymerase chain	pressure differential
positioning system	retail environment
programmable logic	risk management
scripting languages	search criteria
semiconductor devices	service offerings
temperature sensor	software engineering
user interface	transitory computer
vapor deposition	uninterruptible power
wireless technology	window assembly

Notes: This table lists examples of bigrams that pass (technologies) or fail (non-technologies) the Wikipedia filter. All examples are among the top 100 bigrams in terms of citation-weighted patent counts or number of job postings. Failure examples are sampled to reflect different sources of rejection in the Wikipedia filter.

Appendix Table 2 – Technology excerpts from earnings calls

Company	EC month	Excerpt
Ambarella Inc	4/2018	results that are many times higher in terms of processing performance per watt In March we successfully demonstrated to customer and investors our fully| AUTONOMOUS VEHICLE or embedded vehicle autonomy on Silicon Valley Road EVA navigated various traffic scenarios presented by Silicon Valleys challenging urban environment The fully autonomous
Cloudera Inc	4/2019	combined company road map which we rolled out in March of this year During this period of uncertainty we saw increased competition from the| PUBLIC CLOUD |vendors Second the announcement in March of Cloudera Data Platform our new hybrid and multicloud offering created significant excitement within our customer base CDP
NVIDIA Corp	7/2015	lot of very exciting development and were working with a lot of them because we have a platform that was really designed to fuse| COMPUTER VISION |cameras from all around the car as well as radars and LIDARS and sonars and be able to do path planning and all of
Proto Labs Inc	1/2015	orders in addition we added capacity to our manufacturing facility in europe in we completed our first acquisition purchasing fineline an| ADDITIVE MANUFACTURING |or 3D printing company based in raleigh north carolina the acquisition was completed last april and is highly complementary to proto labs roughly of our customers use
Cellectar Biosciences Inc	10/2017	collaboration with Acunova Therapeutics each provide these types of strategic benefits Avicenna provides us with the unique opportunity to collaborate with experts in the antibody| DRUG CONJUGATE |or ADC field Not only does this provide the opportunity to work with a very promising small molecule payload but it also allows
L-3 Communications Holdings Inc	10/2002	metal detectors where they always make you take your shoes off This is a passive scanner as I told some of you It uses| MILLIMETER WAVE |It is nonintrusive and causes no harm or disease It will guarantee you won't have a weapon on you of any kind or be
InvenSense Inc	7/2016	as they strive to enable improved locationbased services and mapping user experience A significant opportunity for increasing our mobile content is UltraPrint our ultrasonic| FINGERPRINT SENSOR |I am very pleased to report that we are on track with the development of this gamechanging technology and have successfully passed several technology
SunPower Corp	10/2006	then be able to participate in the global electricity market which is measured in the form of trillion We have direct control over the solar cell and| SOLAR PANEL |portions of the value chain the technology core of the value chain that represents to of total installed costs in these
Donnelley Financial Solutions Inc	4/2018	speed and improve both the quality and consistency of business results for our clients in capital markets through the introduction of| MACHINE LEARNING |and artificial intelligence we will improve the efficiency of XBRL tagging and align with the efforts at the SEC to move from documents to data This investment

Notes: This tables presents examples of earning calls excerpts (in column 3) with 25 words before and after the mention of a technology bigram, with the firm (in column 1) and the date of the earnings call (in column 2).

Appendix Table 3 –List of Wikipedia titles mentioned in >100 earnings calls

Mobile device	Bank account*	Music Player Daemon*	Genetic marker	Electronic discovery
Mobile phone	Solar panel	Laser diode	Combined Charging System	Linux powered device
Adverse event*	Communication channel	Nearfield communication	Vertical launching system	Drug design
User interface	Software defined radio	Power semiconductor device	Low Pin Count	Viral vector
Financial instrument*	Predictive modelling*	Thermography	Ion channel	Growth medium*
Smartphone	Semiconductor device	Heat treating	Plasma display	Sensor fusion
Wireless network	Single instruction multiple data	Performance appraisal*	Computer network	Pattern recognition
Active users*	Domain name*	Microsoft Access*	Electronic stability control	Input device*
Flat panel display	Barcode	Client computing	Wireless sensor network	Video capture
Digital content*	Project delivery method*	Dataspaces	Search algorithm*	Display device
Combination therapy	Distributed antenna system	Leased line*	Active noise control	Gesture recognition
Digital video	Sport utility vehicle	Dental implant	Urine collection device	Caregiver*
Model organism	Payment card	Multifunction printer*	Tunable laser	Policy entrepreneur*
Monoclonal antibody	Minimally invasive procedure	Tablet computer	Fingerprint*	Autonomous system Internet
Data source name*	Extremely high frequency	Semiconductor memory	Channel access method	Emergency Broadcast System*
Keypad	Central processing unit	DNA sequencing	Peritoneal dialysis	Oriented strand board
Social networking service	Location based service	Optical module	Error detection and correction	Variable computer science*
Digital camera	Lithium ion battery	Software versioning	Erotic target location error*	Robotic arm
Wireless LAN	Microsoft Windows*	User profile	Electronic game	Image organizer
Wireless	Unstructured data*	Data loss*	Computer simulation	Barcode reader
Transaction account	Compression seal fitting*	USB flash drive	Transdermal patch	Soulseek
Flash memory	LED circuit	Touchpad	File format*	Optical medium
Data communication	Baseband*	Powder coating	Semiconductor device fabrication*	Epidermal growth factor
Networking hardware	Digital image processing	Rapid prototyping	Dynamic random access memory	Cloud computing
Information privacy*	Video clip*	Continuous glucose monitor	Internet filter	Western blot
Online game	Multilayer perceptron*	Hematopoietic stem cell transplantation	Motion capture	Windowing system
Hard disk drive	Airborne early warning and control	Mental chronometry*	Solar thermal energy	Reduced relative clause*
Video game console	Cell site	Serverless computing*	Speed networking	Email
Mobile computing	Laptop	Liquid crystal display	Atomic layer deposition	
Digital television	Data model*	Silicon germanium	Fluid dynamics	
Media Player Windows*	Network virtualization	Data store	Drug eluting stent	
VLC media player*	File system*	Aggregate data*	Meta analysis*	
Virtual reality	Unmanned aerial vehicle	Ecommerce	Channel blocker	
Immortalized cell line	Portable media player	Data type*	Typical versus maximum performance*	
Personal computer	Multicore processor	Google Search*	Flexible electronics	
Nature based solutions*	Effective dose radiation*	Light emitting diode	SMS	
Selective Service System*	Web feed	Server computing	Video editing	
Network operating system	LED lamp	Component based software engineering	Digital signature	
Digital imaging	Facial recognition system	Software distribution	Electrooptical sensor	
Programming tool	Printing	Microcontroller	Identity verification service	
Data quality*	Antibody drug conjugate	Fiberoptic cable	Selective serotonin reuptake inhibitor	

Self driving car	Insulin pump	List of search engines	CMOS
Passive optical network	Voicemail	Push technology	Optical disc drive
Augmented reality	Speech recognition	Machine learning	Magnetic resonance imaging
Smart card	Satellite navigation device	Data file*	Amazon Relational Database Service
Card security code	Building automation	Thermometer	Programming language*
Power electronics	User activity monitoring	Software architecture	Computer data storage
Text messaging	Data integrity	Dendritic cell*	Medical imaging
Home automation	Programmable logic device	Therapeutic drug monitoring	Ventricular assist device
Receptor antagonist	Web query	Genetic engineering	Dry powder inhaler
Instant messaging	Flow cytometry	Da Vinci Surgical System*	Network switch
Multimedia Messaging Service	Virtual private network	Mesh networking	Urine test strip
Coiled tubing	Desktop computer	Epidermal growth factor receptor	Label printer
Digital subscriber line	Network monitoring	Monoclonal antibody therapy	Distributed computing
Systems architecture*	Audio file format*	Communication protocol*	Physical media
OS level virtualization*	Lithium battery	Neural network	Glucose meter
Software Updater	Global Positioning System	Docking station	Electronic program guide
Patch computing	Visitor Based Network	Recombinant DNA	Adeno associated virus
Expected value*	Active site*	Duty cycle*	List of copy protection schemes*
Agent based model	Memory card	Mass spectrometry	Debit card
Data access object*	Gallium nitride	Data compression*	Laser scanning
Hybrid electric vehicle	Load balancing computing	Cruise control	Intensive care unit*

Notes: This table lists all technologies (Wikipedia titles) associated with technology bigrams which are mentioned in at least 100 earnings calls. * next to a technology indicates that it is dropped in human audit described in Section 7.

Appendix Table 4 – Top technologies by pioneer states, by share of early patenting

Region	Division	State Name	Title	Pct. Early Patents	Year of Emergence
(1)	(2)	(3)	(4)	(5)	(6)
Northeast	New England	Connecticut	Label printer	6.19	1986
		Massachusetts	Antibody-drug conjugate	13.58	1998
		New Hampshire	Computer data storage	14.3	1992
	Mid Atlantic	New Jersey	Transaction account	19.45	2000
		New York	Digital imaging	16.95	1992
		Pennsylvania	Machine learning	17.12	1992
Midwest	East North	Illinois	Ventricular assist device	12.59	1983
		Indiana	Hybrid electric vehicle	8.61	1995
		Michigan	Electronic stability control	49.93	1996
		Ohio	Da Vinci Surgical System	17.5	2005
		Wisconsin	Peritoneal dialysis	15.55	1990
	West North	Minnesota	Optical disc drive	14.15	1991
		Missouri	Sensor fusion	24.64	1996
South	South Atlantic	Florida	Dental implant	12.87	1987
		Georgia	Oriented strand board	8.34	1991
		Maryland	Adeno-associated virus	13.28	1994
		North Carolina	Gallium nitride	24.36	1993
		Virginia	Selective Service System	7.25	1990
	East South	Tennessee	Lithium battery	4.79	1986
	West South	Texas	Coiled tubing	51.74	1988
West	Mountain	Arizona	Cruise control	4.71	1993
		Colorado	Network monitoring	23.1	1991
		Idaho	Dynamic random-access memory	10.11	1991
		Utah	Fluid dynamics	5.62	1993
	Pacific	California	Continuous glucose monitor	73.55	1996
		Oregon	Multifunction printer	29.59	1997
		Washington	Distributed computing	25.99	1987

Notes: This table reports top technology (in column 4) by share of early patenting for a sample of states (in column 3). We report the corresponding census division in column 2, and the census region in column 1. The sample of states are those which file at least 50,000 patents during our sample period. In column 6, we report the share of early patenting, calculated as $Share\ Early\ Patenting_{\tau,s} = \frac{Early\ Patents_{\tau,s}}{\sum_j Early\ Patents_{\tau,j}}$ for every technology τ and state s. We drop the titles "policy entrepreneur," "caregiver," and "reduced relative clause" that pass the Wikipedia criteria of technology bigrams but are removed in the human audit of technology bigrams.

Appendix Table 5 — Share of patenting and robustness

Number of Earnings Calls Mentions	Silicon Valley, Boston, New York, Seattle	Silicon Valley	Unweighted	Number of Job Postings	Bins by Job Postings
	(1)	(2)	(3)		(4)
+500	46.490***	26.745***	44.712***	+1,000k	49.465***
	(1.465)	(1.607)	(1.400)		(2.161)
100 to 499	37.344***	21.409***	39.688***	500k to 1,000k	49.669***
	(1.369)	(1.095)	(1.040)		(2.494)
40 to 99	31.933***	17.618***	33.335***	200k to 500k	43.557***
	(1.413)	(1.171)	(1.013)		(1.903)
10 to 39	30.402***	14.664***	35.384***	50k to 200k	38.643***
	(0.847)	(1.005)	(0.973)		(1.856)
3 to 9	27.240***	13.002***	33.472***	10k to 50k	40.487***
	(1.273)	(1.127)	(1.012)		(2.010)
0 to 2	24.607***	10.167***	30.941***	0 to 10k	32.979***
	(0.708)	(0.806)	(0.901)		(1.226)

Notes: Column 1 shows the result from a regression of concentration at the bigram-level on indicators for the number of earnings calls that mention a given bigram. For each bigram, concentration is measured by the share of patenting associated with that bigram in the top five CBSAs (San Jose-Sunnyvale-Santa Clara, CA; San Francisco-Oakland-Hayward, CA; New York-Newark-Jersey City, NY-NJ-PA; Seattle-Tacoma-Bellevue, WA; Boston-Cambridge-Newton, MA-NH). The top five CBSAs are the five regions with the highest number of patenting associated with technologies with more than 100 earnings call mentions. Observations are weighted by the total number of patents associated with the given bigram. Column 2 measures concentration as the share of patenting in the top two CBSAs, measured similarly (CBSAs: San Jose-Sunnyvale-Santa Clara, CA and San Francisco-Oakland-Hayward, CA). In column 3, we use the top five CBSAs, but with equal weights across observations. Column 4 reproduces the regression but defines the bins in terms of job postings. A joint F-test of the equality of the six coefficients was rejected at the 0.01 confidence level in all regressions.

Appendix Table 6 – Top occupations for technology - “Machine Learning”

Occupations	Pct. “Machine Learning” postings	Total Job Postings
Computer and Information Research Scientists	40.83	179,636
Astronomers	3.33	10,236
Life Scientists, All Other	2.41	27,691
Computer Hardware Engineers	2.38	91,259
Statisticians	2.31	193,339
Computer Science Teachers, Postsecondary	2.24	34,221
Operations Research Analysts	1.80	874,226
Database Administrators	1.62	1,110,414
Social Science Research Assistants	1.56	53,040
Biological Scientists, All Other	1.55	97,722
Software Developers, Applications	1.50	6,963,792
Physical Scientists, All Other	1.21	12,176
Engineering Teachers, Postsecondary	1.17	13,606
Social Scientists and Related Workers, All Other	1.15	59,306
Detectives and Criminal Investigators	1.11	130,801
Biomedical Engineers	1.06	17,543
Computer and Information Systems Managers	1.04	205,267
Financial Specialists, All Other	1.03	290,386
Architectural and Engineering Managers	1.01	586,301
Computer Occupations, All Other	1.00	6,172,457

Notes: This table lists the top occupations (in column 1) by the share of postings that are associated with technology “Machine Learning” (in percent, in column 2). Column 3 reports the total job postings for the occupation. The technology “Machine Learning” refers to Wikipedia title “Machine learning” and the associated bigrams “machine learning” and “learning algorithms.”

Appendix Table 7 – Human audit results of technology job postings

Panel A: Audit results			
Audit	Use	Produce	Total
Describes company	6%	10%	16%
Describes task	46%	34%	80%
Neither	NA	NA	4%
Panel B: Audit results after clipping top 50 and bottom 50 words			
Audit	Use	Produce	Total
Describes company	2%	2%	4%
Describes task	55%	36%	91%
Neither	NA	NA	5%
Panel C: Examples excerpts			
Produce	"we are looking for a ux developer to join our fast growing team our mission is to make your interactions with **touchscreens** more interesting more natural and more engaging we have developed a novel haptic **touchscreen** technology that not only tracks the fingertips but controls what they feel"		
Use	"duties of this job the employee is required to occasionally use clarity of vision at approximately feet or more stoop bend the body downward and forward by the spine at the waist frequently use a keyboard, key **touch screen,** or mouse to enter text or data into a computer"		
Neither	"our super cool office space which doesn't feel like an office is designed with our employees in mind techy surroundings a great outdoor space with Wi-Fi hookups for your laptop plus Bluetooth capabilities for **music streaming** we enjoy cultivating a supportive and all around positive culture that keeps our employees happy this will be a place you will want to come to everyday"		

Notes: This table presents the results from a human audit of Burning Glass technology job postings. As a part of the human audit, we classify each of 1,000 randomly sampled job postings into two ways: 1) whether the technology reference in the job posting refers to the company or the task content of the job posting, and 2) whether the job describes the use or the production of the technology. See the main text for details. In Panel A, we perform the audit on the original text of job postings for a smaller set of 100 postings. In Panel B, we clip the text of job postings by 50 words at the top and bottom, resample 1,000 postings, and then repeat the audit. Panel C provides examples of job postings classified under the "use," "produce," and "neither" categories.

Appendix Table 8 – Summary statistics

Variable	(1) N	(2) Mean	(3) SD	(4) p25	(5) p50	(6) p75
Panel A: Technology bigrams						
Postings	1,899	26633.79	127987.98	45.00	534.00	6440.00
Earnings Calls	1,899	132.86	471.41	1.00	13.00	85.00
Cite wt. Patents	1,899	4643.53	7289.85	1506.59	2513.45	5140.56
Emergence Year	1,899	1991.54	5.28	1989	1991	1994
# Pioneers	1,899	5.90	2.60	4	6	7
Panel B: Technology bigram by year						
Tech Postings	8,347	6,233	20,736	236	764	3,462
Coefficient of Variation	8,347	4.69	3.10	2.36	3.77	6.29
Share College Educated	8,347	52.30	11.26	45.50	53.99	59.53
Wage	8,347	59,846	9,938	53,043	61,452	66,757
Share Research Postings	8,347	6.26	6.70	1.73	4.31	8.61
Share Produce Postings	8,347	7.08	8.06	2.13	5.06	9.02
Share RDP Postings	8,347	11.56	10.40	4.32	8.97	15.77
Share Training Postings	8,347	29.25	19.12	13.23	27.46	42.04
Panel C: CBSA by technology bigram by year						
Total Postings	17,634,114	20670.29	81760.65	1252	2825	8897
Tech Postings	17,634,114	2.96	67.09	0.00	0.00	0.00
Normalized Share	14,391,504	0.58	2.77	0.00	0.00	0.28

Notes: This table shows summary statistics – the number of observations, mean, standard deviation, 25th percentile, median and 75th percentile - for variables used in the analyses of the paper. Panel A reports statistics for our sample of technology bigrams. Panel B reports summary statistics for our datasets at the technology bigram x year level, which are used in region broadening (Table 4) and skill broadening (Table 7) regressions. Panel C reports summary statistics for our CBSA x technology bigram x year data used in pioneer advantage regressions in Table 5. The normalized share of bigram jobs in all panels is calculated as $Normalized\ share_{c,\tau,t} = \frac{share\ jobs\ exposed_{c,\tau,t}}{share\ jobs\ exposed_{\tau,t}}$, where c is a location (CBSA), τ is a technology, and t is calendar year. The coefficient of variation is calculated using the normalized share of technology job postings by dividing the standard deviation of $Normalized\ share_{c,\tau,t}$ across locations c in year t by its mean in year t for each technology bigram τ. Skill level variables (in Panel B – Share College Educated and Wage) are calculated using $Skill_t^{\tau} = \frac{\sum_o N_{o;t}^{\tau} \chi_{o;2015}}{\sum_o N_{o;t}^{\tau}}$, where $N_{o,t}^{\tau}$ is the number of Burning Glass job postings mentioning technology bigram τ that are in SOC code o at time t, and $\chi_{o,2015}$ is the average skill level for occupation o, as measured by the 2015 ACS. The share of research/produce/RDP/training postings are calculated for each technology bigram τ and year t according to the process explained in Section 5a.

Appendix Table 9 – Robustness: Testing for non-linearity in `region broadening'

Panel A: Comparing linear, quadratic and logarithmic specifications

	$Coefficient\ of\ Variation_{\tau,t}$			log ($Coefficient\ of\ Variation_{\tau,t}$)
	(1)	(2)	(3)	(4)
$Years\ since\ emergence_{\tau,t}$	-0.068***	-0.028	-0.305***	-0.015**
	(0.026)	(0.133)	(0.076)	(0.006)
$(Years\ since\ emergence_{\tau,t})^2$		-0.001	0.004**	
		(0.003)	(0.002)	
Constant (CV at $t_{\tau,0}$)	5.577***	5.203***	9.008***	1.523***
	(0.645)	(1.444)	(0.861)	(0.158)
R-squared	0.019	0.019	0.817	0.017
N	4,270	4,270	4,270	4,270
Bigram FE	NO	NO	YES	NO

Panel B: Testing for general non-linearity

Specification: $Coefficient\ of\ Variation_{\tau,t} = \beta_0 + \beta_1 * (years\ since\ emg_{\tau,t})^{\beta_2}$

Initial conditions: $\beta_0 = mean(Coefficient\ of\ Variation_{\tau,t}), \beta_2 = 1$

β_0	β_1	β_2
5.134***	-0.009	1.536**
(0.375)	(0.020)	(0.626)
H0: $\beta_2 = 1$	0.73	
p-value	0.392	
H0: $\beta_1 = \beta_2 = 0$	1,895.00	
p-value	0.000	

Notes: This table reports results from tests for non-linearity of the relationship between the $Coefficient\ of\ Variation_{\tau,t}$ and $Years\ since\ Emergence_{\tau,t}$. Panel A, column 1 reports results from a regression of the coefficient of variation on the year since emergence, the same specification as in Table 4, Panel A, column 1. Column 2 adds a quadratic term to this regression and column 3 adds technology bigram fixed effects. Column 4 presents results for a regression where we replace $Coefficient\ of\ Variation_{\tau,t}$ with log ($Coefficient\ of\ Variation_{\tau,t}$) as the dependent variable. Panel B tests for non-linearity by estimating the specification: $Coefficient\ of\ Variation_{\tau,t} = \beta_0 + \beta_1 * (years\ since\ emg_{\tau,t})^{\beta_2}$ with non-linear least squares and initial conditions $\beta_0 = mean(Coefficient\ of\ Variation_{\tau,t})$ and $\beta_2 = 1$. The regressions are restricted to the sample of technology bigrams that appear in at least 100 earnings calls. For more details on the specification, refer to the note in Table 4.

Appendix Table 10 – Robustness: Pioneer location advantage in technology hiring

	$Normalized\ Share_{c,t,\tau}$			
	(1)	(2)	(3)	(4)
$Rate\ of\ decline\ per\ year$	-0.027***	--0.027***	--0.024***	--0.024***
	0.003	0.003	0.005	0.005
Implied years to zero advantage	37.04	37.04	41.67	41.67
$Pioneer_{c,\tau}$	1.321***	1.323***	1.118***	1.121***
	(0.254)	(0.255)	(0.294)	(0.295)
$Pioneer_{c,\tau} * Years\ since\ emg_{\tau,t}$	-0.035***	-0.035***	-0.027**	-0.027**
	(0.011)	(0.011)	(0.012)	(0.012)
R-squared	0.030	0.037	0.033	0.040
N	7,751,024	7,751,024	7,751,024	7,751,024
Bigrams	835	835	835	835
Bigram FE	YES	NA	YES	NA
CBSA FE	YES	YES	NA	NA
Year FE	YES	NA	YES	NA
Bigram x Year FE	NO	YES	NO	YES
CBSA x Year FE	NO	NO	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title	Wiki Title

Notes: This table reports robustness checks for results from the pioneer advantage regressions in Table 5. We are focusing on the change in the rate of decline per year, so place these findings above the regression coefficients. Column 1 repeats the regression specification from Table 5, column 3. Columns 2, 3, and 4 add CBSA x Bigram fixed effects and CBSA x Year fixed effects iteratively. The first row reports the decay rate, as defined in Table 5. The regressions use all technology bigrams that appear in at least 1000 job postings in our sample. For more details on the specification, refer to the note in Table 5.

Appendix Table 11 – Top occupations for example technologies

Emergence Year	Technology	Top Occupations (by pct. share of postings)
1979	Hard disk drive	Broadcast Technicians (4.5); Computer Operators (2.7)
1980	Barcode reader	Packers and Packagers, Hand (0.5); Library Assistants, Clerical (0.2)
1981	Laser diode	Chiropractors (0.4); Physicists (0.3)
1982	Personal computer	Automotive Glass Installers and Repairers (26.7); Demonstrators and Product Promoters (22.6)
1983	Flatpanel display	Upholsterers (2.6); Audio and Video Equipment Technicians (2.0)
1984	User interface	Multimedia Artists and Animators (26.7); Web Developers (23.4)
1985	Mobile phone	Graders and Sorters, Agricultural Products (20.7); Advertising Sales Agents (15.8)
1986	Facial recognition system	Marriage and Family Therapists (1.8); Child, Family, and School Social Workers (1.6)
1987	Digital video	Audio and Video Equipment Technicians (7.4); Film and Video Editors (5.1)
1988	Model organism	Astronomers (2.3); Life Scientists, All Other (1.8)
1989	Mobile device	Electronic Equipment Installers and Repairers, Motor Vehicles (12.6); Foresters (11.3)
1990	Debit card	Pharmacy Aides (10.3); Railroad Conductors and Yardmasters (5.5)
1991	Flash memory	Computer Hardware Engineers (0.9); Radio, Cellular, and Tower Equipment Installers and Repairs (0.5)
1992	Machine learning	Computer and Information Research Scientists (56.7); Astronomers (4.0)
1993	Financial instrument	Financial Specialists, All Other (0.7); Economists (0.4)
1994	Active users	Advertising Sales Agents (2.0); Sales Representatives, Services, All Other (0.3)
1995	Hybrid electric vehicle	Electronics Engineers, Except Computer (0.2); Electrical Engineers (0.2)
1996	Digital content	Producers and Directors (3.6); Multimedia Artists and Animators (3.6)
1997	Multicore processor	Computer Hardware Engineers (0.2); Electronics Engineers, Except Computer (0.1)
1998	Information privacy	Information Security Analysts (2.7); Financial Specialists, All Other (0.7)
1999	Unmanned aerial vehicle	Avionics Technicians (2.5); Commercial Pilots (1.4)
2000	Transaction account	Traffic Technicians (0.2); Brokerage Clerks (0.2)
2001	Smartphone	Automotive Glass Installers and Repairers (28.5); Home Appliance Repairers (21.5)
2002	Online game	Fine Artists, Including Painters, Sculptors, and Illustrators (1.2); Multimedia Artists and Animators (0.8)
2003	Social networking service	Reporters and Correspondents (5.4); Radio and Television Announcers (3.4)
2004	Electronic discovery	Graders and Sorters, Agricultural Products (13.2); Parts Salespersons (1.3)
2005	LED circuit	Electrical and Electronics Drafters (0.2); Electronics Engineers, Except Computer (0.1)
2006	Augmented reality	Computer Hardware Engineers (0.3); Interior Designers (0.3)
2007	Self-driving car	Computer Hardware Engineers (0.5); Armored Assault Vehicle Crew Members (0.2)

Notes: This table lists the top two occupations and the percentage of postings (in column 3) that mention technology bigrams associated with top technologies (in column 2). The list of technologies is the same as our list in Table 1.

Appendix Table 12 – Robustness: Region broadening and pioneer advantage by skill

Panel A: Region Broadening			
	$\log(Coefficient\ of\ Variation)_{\tau,t}$		
	(1)	(2)	(3)
Skill Level	High	Medium	Low
$Years\ since\ emergence_{\tau,t}$	-0.038***	-0.033***	-0.027***
	(0.003)	(0.003)	(0.002)
R-squared	0.845	0.844	0.837
N	8,069	8,069	8,069
Bigrams	835	835	835
Bigram FE	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title

Panel B: Pioneer Advantage			
	$Normalized\ Share_{c,\tau,t}$		
	(1)	(2)	(3)
Skill Level	High	Medium	Low
$Pioneer_{c,\tau}$	1.319***	0.921***	1.112***
	(0.233)	(0.225)	(0.253)
$Pioneer_{c,\tau} * year\ since\ emg_{\tau,t}$	-0.029***	-0.022**	-0.036***
	(0.010)	(0.010)	(0.010)
R-squared	0.016	0.018	0.012
N	8,581,946	8,567,285	8,395,144
Bigrams	835	835	835
Bigram FE	YES	YES	YES
Std. Errors (cluster)	Wiki Title	Wiki Title	Wiki Title
Rate of decline per year	**-0.022**	**-0.024**	**-0.032**
	(0.004)	**(0.005)**	**(0.003)**

Notes: This table reports region-broadening regressions for low-, medium-, and high-skill job postings at the technology bigram x year level. Columns 1, 2 and 3 show separate regressions for high-skill (column 1), medium-skill (column 2), and high-skill (column 3) job postings. For definitions of these concepts see Section 4.c of the main text. The regressions are restricted to the sample of technology bigrams that appear in at least 1000 job postings. Standard errors are clustered by Wikipedia title.

Appendix Table 13 – Keywords for research, development, and production (RDP) job postings

Keyword	Pct, Tech. Job Postings	Keyword	Pct, Tech. Job Postings
design and*	2.81	in creating	0.1
research*	1.54	the leading	0.072
development of*	1.049	the worlds	0.056
to create*	0.555	to market	0.052
and develop*	0.545	adoption of	0.033
range of	0.456	product engineering	0.018
designing and*	0.45	customization of*	0.015
developing and*	0.38	to customize*	0.01
in developing*	0.344	enhancements of	0.005
leader in	0.315	and experimentation*	0.005
research and*	0.31	and exploit*	0.004
and developing*	0.307	exploration of*	0.003
high performance	0.274	exploitation of*	0.003
build and	0.255	to personalize	0.002
team of	0.236	in expanding	0.002
a global	0.209	and commercializing	0.001
product development	0.207		
in building	0.147		

Notes: This table reports the list of keywords used to classify technology job postings into those requiring research, development, and production (RDP) of the technology. We flag job postings that mention these RDP keywords within 15 words of a technology bigram as requiring the RDP of the technology. The asterisk * identifies keywords for research and development.

Appendix Table 14 – Robustness: Skill broadening robustness

	Share College Educated			
	Baseline (2010-2019)	W/ 2007	2010-2015	2016-2019
	(1)	(2)	(3)	(4)
Year since emergence	-0.288***	-0.294***	-0.260***	-0.228**
	(0.079)	(0.072)	(0.088)	(0.114)
Baseline measure	59.095***	59.313***	58.901***	57.047***
	(1.794)	(1.605)	(1.814)	(2.933)
R-squared	0.019	0.021	0.016	0.007
N	8,347	9,277	5,008	3,339
Bigram FE	NO	NO	NO	NO

Notes: This table reports robustness of skill broadening results. All regressions are at the technology bigram x year level. The dependent variable is the average share of job postings mentioning technology bigram τ in year t that require a college degree. Column 1 uses our baseline sample of Burning Glass job postings from 2010-2019 (replicating Table 7, Panel A, column 2); column 2 adds 2007 to our baseline sample. Columns 3 and 4 split the sample into two: column 3 only includes 2010-2015 and column 4 only includes 2016-2019. The regressions use all technology bigrams that appear in at least 1000 job postings in our sample. For more details on the specification, refer to the note in Table 7.

Appendix Table 15 – Keywords for "training" job postings

Keyword (1)	Pct, Tech. Job Postings (2)	Keyword (1)	Pct, Tech. Job Postings (2)
knowledge of	9.491	understanding in	0.034
experience with	8.311	proficiency of	0.019
experience in	5.617	expertise of	0.018
understanding of	4.075	specialization in	0.014
working knowledge	2.112	experiences with	0.014
responsible for	2.031	competence of	0.003
familiarity with	1.987	indepth knowledge	0.002
experience using	1.242	credentials in	0.001
proficiency in	1.168	understanding in	0.034
proficient in	0.963		
skills in	0.757		
expertise in	0.694		
familiar with	0.61		
knowledge in	0.527		
experience of	0.249		
extensive knowledge	0.234		
knowledge with	0.12		
grasp of	0.057		

Notes: This table reports (in column 1) the list of keywords used to classify technology job postings into those that require training in the technology. We flag job postings that mention these training keywords within 15 words of a technology bigram as requiring training or experience with that technology. Column 2 reports the respective share of technology job postings in percent which require training.

Appendix Table 16 – Top industries by technology

Emergence Year (1)	Technology (2)	Top Pioneer Industry (3)
1979	Hard disk drive	Computer and Peripheral Equipment Manufacturing
1980	Barcode reader	Computer and Peripheral Equipment Manufacturing
1981	Laser diode	Computer and Peripheral Equipment Manufacturing
1982	Personal computer	Computer and Peripheral Equipment Manufacturing
1983	Flatpanel display	Semiconductor and Other Electronic Component Manufacturing
1984	User interface	Computer and Peripheral Equipment Manufacturing
1985	Mobile phone	Communications Equipment Manufacturing
1986	Facial recognition system	Software Publishers
1987	Digital video	Computer and Peripheral Equipment Manufacturing
1988	Model organism	Pharmaceutical and Medicine Manufacturing
1989	Mobile device	Computer and Peripheral Equipment Manufacturing
1990	Debit card	Management, Scientific, and Technical Consulting Services
1991	Flash memory	Computer and Peripheral Equipment Manufacturing
1992	Machine learning	Software Publishers
1993	Financial instrument	Computer and Peripheral Equipment Manufacturing
1994	Active users	Management, Scientific, and Technical Consulting Services
1995	Hybrid electric vehicle	Motor Vehicle Manufacturing
1996	Digital content	Software Publishers
1997	Multicore processor	Semiconductor and Other Electronic Component Manufacturing
1998	Information privacy	Software Publishers
1999	Unmanned aerial vehicle	Aerospace Product and Parts Manufacturing
2000	Transaction account	Non-depository Credit Intermediation
2001	Smartphone	Information Services
2002	Online game	Software Publishers
2003	Social networking service	Other Information Services
2004	Electronic discovery	Pharmaceutical and Medicine Manufacturing
2005	LED circuit	Semiconductor and Other Electronic Component Manufacturing
2006	Augmented reality	Semiconductor and Other Electronic Component Manufacturing
2007	Self-driving car	Other Information Services

Notes: This table lists the top pioneer industry (NAICS four-digit codes) (in column 3) by percentage of job postings associated with a sample of technologies (in column 2) corresponding to emergence years (in column 1). The list of technologies is the same as in Table 1.

Appendix Table 17 – Wikipedia titles and trigrams

Wikipedia Title (1)	Trigrams (2)	Wikipedia Title (1)	Trigrams (2)
Web-RTC	real time communications	DNA-binding domain	dna binding domain
Real-time operating system	time operating system; real time operating	LTE telecommunication	term evolution lte
Injection molding machine	injection molding machine	Perchloric acid	sulfuric acid nitric
MEMS	electro mechanical systems; electro mechanical system	Phase detector	phase frequency detector
Single-mode optical fiber	single mode fiber; single mode fibers	Vinyl-sulfonic acid	meth acrylic acid
Heat recovery steam generator	waste gas stream; heat recovery steam	Water-gas shift reaction	water gas shift
Mobile phones on aircraft	portable electronic devices	Calcium oxide	calcium oxide calcium
Optical fiber connector	optical fiber connector; fiber optic connectors; fiber optic connector	Carboxylic acid	alkyl alkenyl aryl
Colony-stimulating factor	colony stimulating factors	Cartesian coordinate system	cartesian coordinate system
Granulocyte colony-stimulating factor	colony stimulating factor	Cellulose acetate phthalate	cellulose acetate phthalate
Protein sequencing	amino acid sequences; terminal amino acid; terminal amino acids	Complement-dependent cytotoxicity	dependent cytotoxicity cdc
Carbon-fiber-reinforced polymers	carbon fiber reinforced	Dent corn	hybrid corn variety
Transaction processing system	transaction processing system	Dibenzylideneacetone	pale yellow solid
History of mobile phones	mobile communication devices	Dimethyl sulfoxide	dimethyl sulfoxide dmso
Engine control unit	engine control unit	Fluorinated ethylene propylene	fluorinated ethylene propylene
Hollow fiber membrane	hollow fiber membrane; hollow fiber membranes	Glyceraldehyde 3phosphate dehydrogenase	glyceraldehyde phosphate dehydrogenase
Satellite navigation	navigation satellite system	Halide	fluoride chloride bromide
Polyvinyl alcohol	poly vinyl alcohol	Ion chromatography	cation exchange chromatography
Electronic control unit	electronic control module	Machine-readable medium and data	machine readable media; computer readable medium; computer readable media; machine readable medium
USB mass storage device class	mass storage devices	Magnesium sulfate	white crystalline solid
Tunnel magnetoresistance	magnetic tunnel junction	Nitrobenzene	pale yellow oil
Ultra-high-molecular-weight polyethylene	molecular weight polyethylene	Phase-shift keying	phase shift keying
Polymethyl methacrylate	poly methyl methacrylate	Phase-transfer catalyst	phase transfer catalyst
Amino acid	amino acid residues	Plant tissue culture	plant cell tissue
Attribute-based access control	access control policy	Sperm-mediated gene transfer	mediated gene transfer
Protein kinase inhibitor	protein kinase inhibitors	Zinc selenide	light yellow solid
Carboxymethyl cellulose	sodium carboxymethyl cellulose		

Notes: This table reports technology trigrams (in column 2) and associated technologies/Wikipedia titles (in column 1). From this list, we exclude trigrams that contain one of the baseline technology bigrams or that correspond to a Wikipedia title (technology) already associated with a technology bigram. Entries are sorted by mentions in earnings calls.

Appendix Table 18 - Wikipedia titles and unigrams

Wikipedia title (1)	Unigrams (2)	Wikipedia title (1)	Unigrams (2)	Wikipedia title (1)	Unigrams (2)
Website	websites	Tomography	tomography	Biomass	biomass
HTTPS	https	Flash ADC	adcs	Multiplexer	multiplexer
Internets	internets	RTP payload formats	rtp	Nonlinear system	nonlinear
Infrastructure	infrastructures	USB	usb	Multiplexing	multiplexes
Video	videos	Biomarker	biomarkers	Playlist	playlists
Upload	uploading ; upload	Defibrillation	defibrillators	Microarray	microarray
PDF	pdf	Pixel	pixels	Transgene	transgenic
Optimizing compiler	optimizations	Apheresis	apheresis	Gac	gac
Blog	blogs	SUV	suvs	Zigbee	zigbee
Middleware	middleware	Bluetooth	bluetooth	Plasmid	plasmid
Intranet	intranets	Solidstate drive	ssds	Waypoint	waypoints
PMOS logic	pmos	Gallium arsenide	gaas	Clear aligners	aligner
Nondeliverable forward	dfs	Pulse oximetry	oximetry	Messenger RNA	mrnas
Prototype	prototyping	Backpack	backpacks	Cannula	cannulas
Endoscopy	endoscopy	Landline	landline	Radiation therapy	radiotherapy
United States Postal Service	usps	PowerPC	rpc	Camcorder	camcorder
Encryption	encrypt ; encryption	Hyperlink	hyperlink	Complex programmable logic device	plds
List of datasets for machinelearning research	datasets	Photonics	photonics	Liquid crystal on silicon	lcos
In vitro fertilization	vfs	Interventional radiology	endovascular	Excipient	excipients
CpG site	cpg	Discrete Fourier transform	dft	OnStar	onstar
Server Message Block	cifs ; smb	Gameplay	gameplay	Histogram	histogram
Polychlorinated biphenyl	pcbs	PCI Express	pcie	Polymerase	polymerases
Pharmacogenomics	genomics	Proteomics	proteomic	Biosensor	biosensor
Billable hours	billable	Integrated circuit	microchips	Hydrocodone	hydrocodone
Topology	topologies	Television timeout	timeout	Oxymorphone	oxymorphone
Clinical decision support system	dsss	Floating production storage and offloading	offloading	TFT LCD	tfts
Interoperability	interoperability	Satellite navigation	gnss	Insulated gate bipolar transistor	igbts
Trusted Platform Module	tpm	Nebulizer	nebulizers	Xenotransplantation	xenograft
Biofuel	biofuel	Biomaterial	biomaterials	Antibody	immunoglobulin
Realtime operating system	rtos	Multimodality	multimodal	Sunscreen	sunscreen

Wikipedia title	Unigrams	Wikipedia title	Unigrams
(1)	(2)	(1)	(2)
Linear particle accelerator	linac	Erlotinib	erlotinib
Rituximab	rituximab ; rituxan	Steam assisted gravity drainage	sagd
Breakpoint	breakpoint	Moisturizer	moisturizing
Embolization	embolization	Supercapacitor	ultracapacitor
Collagen	collagens	Protein isoform	isoforms
Reticle	reticle	Stevia	stevia
Through silicon via	tsvs	Paclitaxel	paclitaxel
Penicillin	penicillins	PSMA scan	psma
Dopamine	dopamine	Lysine	lysines
Duplexer	duplexer	Doxorubicin	doxorubicin
Syringe	prefilled	Prostaglandin	prostaglandins
Megabyte	megabytes	Calcitonin	calcitonin
Small Formfactor			
Pluggable	pluggable	Tamoxifen	tamoxifen
Glycemic index	glycemic	Irinotecan	irinotecan
Naloxone	naloxone	Bupropion	bupropion
Photomask	photomask	Cyclophosphamide	cyclophosphamide
Angiogenesis	angiogenic; angiogenesis	Atorvastatin	atorvastatin
Microparticle	microsphere	Budesonide	budesonide
Biopolymer	biopolymer	Cytarabine	cytarabine
Revascularization	revascularization	Sulfonylurea	sulfonylureas
Joystick	joysticks		
Glycoprotein	glycoprotein		
Backlight	backlights		
Carbonless copy paper	carbonless		
Cephalosporin	cephalosporins		
Polymethyl methacrylate	pmma		
Neurostimulation	neurostimulation		
Chemokine	chemokines		
Cetuximab	erbitux		
Nucleoside	nucleosides		

Notes: This table reports the unigrams (in column 2) and their respective technologies/Wikipedia titles (in column 1), in the order of total job postings that they appear in. The table lists unigrams that appear in at least 100 earnings calls. Out of a total of 200 such unigrams, we exclude 53 unigrams that correspond to existing Wikipedia titles (technologies) already associated with a technology bigram.

Appendix Table 19 – Comparison of main results, by number of earnings call mentions

Panel A: Bigrams with >100 earnings call mentions				
Dependent Variable	Share College Educated	Coefficient of Variation	Normalized Share	Log (Coefficient of Variation)
Result:	Skill Broadening	Region Broadening	Pioneer Persistence	Differential Skill Broadening
	(1)	(2)	(3)	(4)
Year since emergence	**-0.541*** **	**-0.147*** **		-0.033***
	(0.046)	**(0.015)**		(0.003)
Pioneer			1.084***	
			(0.309)	
Pioneer * Year since emg.			-0.032**	
			(0.013)	
Low skill * Year since emg.				**-0.009*** **
				(0.003)
Baseline measure	64.289***	7.400***	0.621***	2.726***
	(1.069)	(0.339)	(0.001)	(0.061)
R-squared	0.895	0.816	0.038	0.852
N	4,270	4,270	3,965,122	12,706
Fixed Effects	Bigram	Bigram	Bigram, CBSA, Year	Bigram
Technologies	428	428	428	428
Rate of decline per year	0.008	0.020	**-0.029 (0.005)**	
Years to Full Decay	62.87	50.43	33.88	.
Panel B: Other bigrams				
Result:	Skill Broadening	Region Broadening	Pioneer Persistence	Region Broadening By Skill
	(1)	(2)	(3)	(4)
Year since emergence	**-0.413*** **	**-0.164*** **		-0.022***
	(0.053)	**(0.017)**		(0.002)
Pioneer			1.986***	
			(0.392)	
Pioneer * Year since emg.			-0.051***	
			(0.015)	
Low skill * Year since emg.				**-0.007** **
				(0.003)
Baseline measure	63.316***	9.977***	0.582***	2.910***
	(1.273)	(0.408)	(0.001)	(0.046)
R-squared	0.919	0.807	0.023	0.796
N	4,307	4,307	3,999,480	12,784
Fixed Effects	Bigram	Bigram	Bigram, CBSA, Year	Bigram
Technologies	431	431	431	431
Rate of decline per year	0.007	-0.016	**-0.026 (0.003)**	.
Years to Full Decay	79.95	60.84	38.94	.
Difference	-0.128(0.070) *	0.017(0.022)	-0.003(0.006)	-0.002(0.005)

Notes: This table reports results from regressions corresponding to our primary results: skill broadening (column 1), region broadening (column 2), pioneer persistence (column 3), and differential region broadening by skill (column 4), separately for those with 100 or more earnings call mentions (Panel A) and other (Panel B) technologies. Our primary coefficients for comparison are in bold in both panels. Panel A is restricted to the sample of technology bigrams that appear in at least 100 earnings calls. Panel B uses all other technology bigrams that appear in at least 1000 job postings in our sample. For more details on the specification, refer to the notes in Tables 4, 5, 6, and 7.

Appendix Figure 1 – Examples of patent text

US007406200B1

(12) **United States Patent**
Syeda-Mahmood et al.

(10) **Patent No.: US 7,406,200 B1**
(45) **Date of Patent: Jul. 29, 2008**

(54) **METHOD AND SYSTEM FOR FINDING STRUCTURES IN MULTI-DIMENSIONAL SPACES USING IMAGE-GUIDED CLUSTERING**

(75) Inventors: **Tanveer Syeda-Mahmood**, Cupertino, CA (US); **Peter J. Haas**, San Jose, CA (US); **John M. Lake**, Cary, NC (US); **Guy M. Lohman**, San Jose, CA (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **11/970,946**

(22) Filed: **Jan. 8, 2008**

(51) **Int. Cl.**

G06K 9/62	(2006.01)
G06F 7/00	(2006.01)
G06F 17/30	(2006.01)
G06F 17/00	(2006.01)

(52) **U.S. Cl.** **382/225**; 707/3; 707/104.1

(58) **Field of Classification Search** 382/224–228; 707/3–6, 104.1

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2004/0013305 A1 1/2004 Brandt et al.
2004/0202370 A1 10/2004 Fisher et al.
2007/0185946 A1 * 8/2007 Basri et al. 708/200

OTHER PUBLICATIONS

Gdalyhu et al., "Self-Organization in Vision: Stochastic Clustering for Image Segmentation, Perceptual Grouping, and Image Database Organization", IEEE Trans. on Pattern Analysis and Machine Intelligence, Oct. 2001, pp. 1053-1074, vol. 23, No. 10.

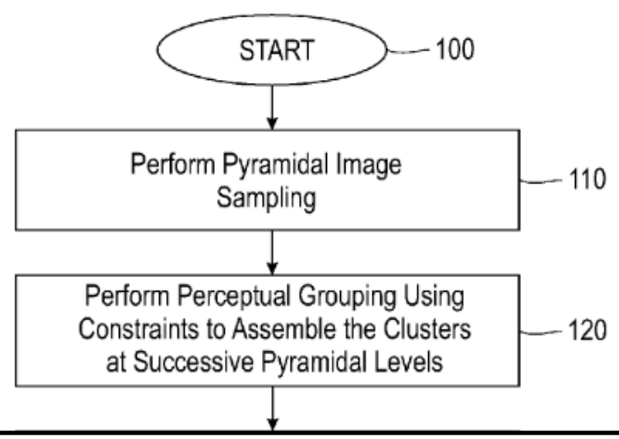
* cited by examiner

Primary Examiner—Aaron W Carter
(74) *Attorney, Agent, or Firm*—Cantor Colburn LLP; Van Ngyen

(57) **ABSTRACT**

A method is provided clustering data points in a multidimensional dataset in a multidimensional image space that comprises generating a multidimensional image from the multidimensional dataset; generating a pyramid of multidimensional images having varying resolution levels by successively performing a pyramidal sub-sampling of the multidimensional image; identifying data clusters at each resolution level of the pyramid by applying a set of perceptual grouping constraints; and determining levels of a clustering hierarchy by identifying each salient bend in a variation curve of a magnitude of identified data clusters as a function of pyramid resolution level.

5 Claims, 10 Drawing Sheets



METHOD AND SYSTEM FOR FINDING STRUCTURES IN MULTI-DIMENSIONAL SPACES USING IMAGE-GUIDED CLUSTERING

CROSS REFERENCE TO RELATED APPLICATIONS

The present application is co-pending with the concurrently filed application, entitled "METHOD AND SYSTEM FOR DATA CLASSIFICATION BY KERNEL DENSITY SHAPE INTERPOLATION OF CLUSTERS," assigned to the assignee of the present application, the contents of which are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to data clustering, and more particularly, to the clustering of multidimensional data to determine high-level structures.

2. Description of Background

Data clustering (or just clustering) is the categorization of objects into different groups, or more precisely, the organizing of a collection of data into clusters, or subsets, based on quantitative information provided by one or more traits or characteristics shared by the data in each cluster. A cluster is a collection of objects which are "similar" between them and "dissimilar" to the objects belonging to other clusters. The goal of clustering is to determine an intrinsic grouping, or structure, in a set of unlabeled data. For example, the functional dependency between two or more time series can lie along a curve. As an example, FIG. **1** shows a graph of a functional dependency between a pair of time series that maps to a perceptible curve having a rotated U-like structure. Clustering can be used to perform statistical data analysis in many fields, including machine learning, data mining, pattern recognition, medical imaging and other image analysis, and bioinformatics.

For applications dealing with sets of high-dimensional data such as multimedia processing applications (for example, content-based image and video retrieval, multimedia browsing, and multimedia transmission over networks), the finding of underlying high-level structures by clustering and categorization is a fundamental analysis operation. A good clustering scheme should, for example, help to provide an efficient organization of content, as well as provide for better retrieval based upon semantic qualities. In video retrieval, because of the larger number of additional features resulting from motion in time, efficient organization is particularly important. In image-based retrieval, semantic quality retrieval is particularly important because clustering provides a means for grouping images into classes that share some common semantics.

Even though clustering of multidimensional datasets is important to determining high-level structures, much of the focus in multidimensional data analysis has been on feature extraction and representation, and existing methods available from data mining and machine learning have been relied on for the clustering task. These methods are primarily based upon the similarity criterion of distance or proximity in which two or more objects belong to the same cluster if they are "close" according to a given distance function that defines a distance between elements of a set (for example, the simple Euclidean distance metric).

The nature of multidimensional datasets, however, presents a number of peculiarities that can lead to misleading or

insufficient results using distance-based clustering, particularly for cases of grouping high-dimensional objects into high-level structures. First, the number of feature dimensions in multidimensional datasets tends to be large in comparison to the number of data samples. As an example, a single four second action video assuming a pair of features per frame (for instance, for representing the motion of the object centroid) can have at least 240 feature dimensions. Similarly, in image clustering, while color, texture, and shape features can encompass hundreds of features, the number of samples available for training could be comparably small. This can result in a data space that is high-dimensional but sparse. The sparseness of the data points can make it difficult to identify the clusters because observation at multiple scales may be needed to spot the patterns.

A second issue that may arise is that the number of clusters for a multidimensional dataset is often unknown and more than one set of clusters may be possible. Different relative scalings can lead to groupings with different structures, even with measurements being taken in the same physical units. To make an informed decision as to relative scaling using existing clustering methods, either the number of clusters needs be known a priori or a hierarchical clustering must be performed that yields several possible clusters without a specific recommendation on one. In a hierarchical clustering, the process builds (agglomerative), or breaks up (divisive), a hierarchy of clusters. The traditional representation of such a hierarchy of clusters is a tree structure called a dendrogram, which depicts the mergers or divisions which have been made at successive levels in the clustering process. A bottom row of leaf nodes represent data and the set of remaining nodes represent the clusters to which the data belong at each successive stage of analysis. The leaf nodes are spaced evenly along the horizontal axis, and the vertical axis gives the distance (or dissimilarity measure) at which any two clusters are joined. Divisive methods begin at the top of the tree, while agglomerative methods begin at the bottom, and cutting the tree at a given height will give a clustering at a selected precision. The bottom level of the hierarchy includes all data points as one cluster.

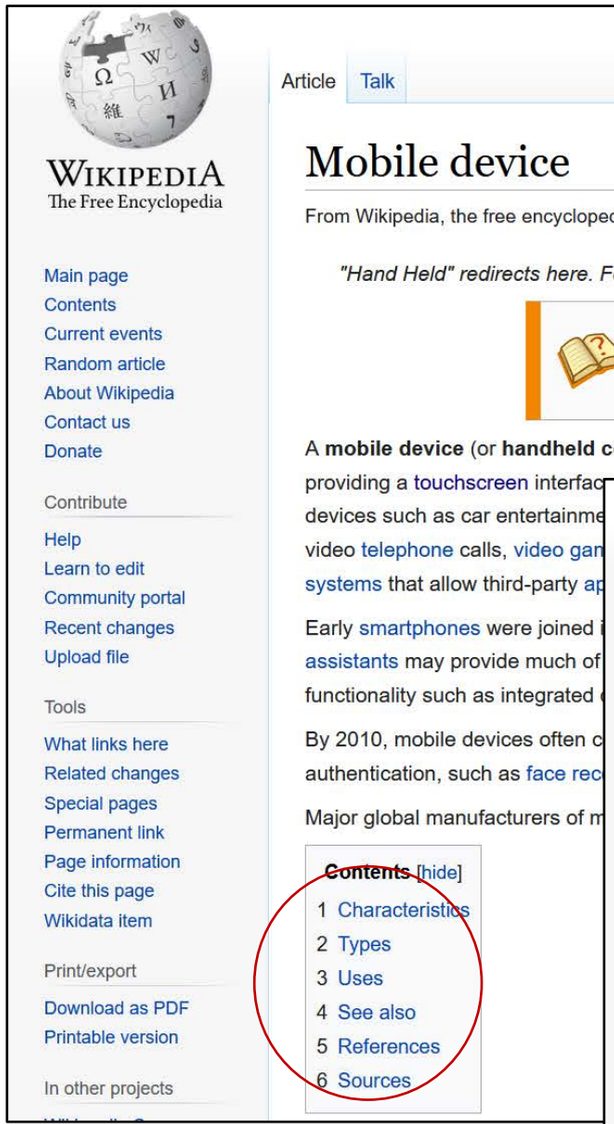
As an example of the scaling issue, a clustering scenario is provided that involves a type of dataset for which the structure of the functional dependency between two or more time series can take a variety of forms. As an example, FIG. **2** illustrates a graph of functional dependencies between a pair of time series in which the noticeable structures are that of three separate lines radiating from common points. While different structures from within this graph may be obtained using hierarchical clustering methods, ideally, it would be desirable to have the result of clustering the dataset indicate the lower level structures (such as the individual splotches in FIG. **2**) as well as the higher-level structures formed (such as the lines perceived in FIG. **2**) without necessarily leading to a single cluster at the top level, unless that is in fact matching how the data collection should be perceived.

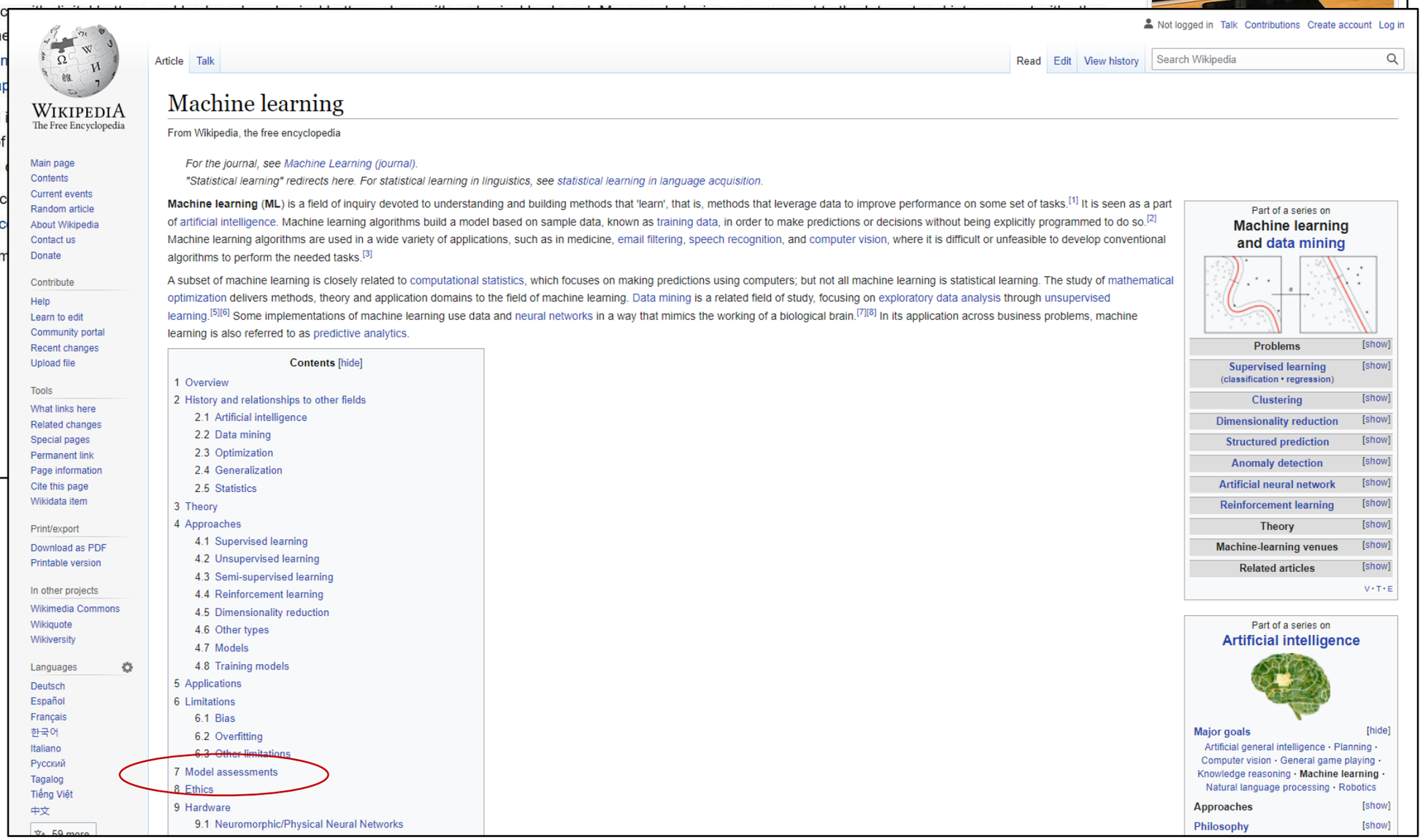
SUMMARY OF THE INVENTION

The shortcomings of the prior art can be overcome and additional advantages can be provided through exemplary embodiments of the present invention that are related to method for clustering data points in a multidimensional dataset in a multidimensional image space. The method comprises generating a multidimensional image from the multidimensional dataset; generating a pyramid of multidimensional images having varying resolution levels by successively performing a pyramidal sub-sampling of the

Notes: These figures show examples of patent text. The left panel displays the first page of U.S. Patent 7406200, while the right panel shows the second page of text in that same patent. For analysis in our paper, we use all text sections, including title, abstract, background, summary, description, and claims.

Appendix Figure 2 – Examples of Wikipedia entries and procedure for filtering technology bigrams





Notes: The figure shows a sample of Wikipedia pages corresponding to two example technologies. The red circles indicate sections that are associated with selection of technology bigrams.

Appendix Figure 3 – Example of a Burning Glass technology job posting, and corresponding location and occupation fields

1.1 CanonCity

Seattle

1.2 CanonCountry

USA

1.3 CanonState

WA

1.4 CleanJobTitle

Server Developer

1.5 JobDate

1/11/2010

1.6 JobDomain

seattle.craigslist.org

1.7 JobText

Date: 2010-01-08, 10:44AM PST

Reply to:

Do you like shaking up the status quo? How about shaking up the advertising industry? Advertising is fundamentally broken in that to manage campaigns across TV, online display, search, and print requires armies of people and has no accountability on the results across the mediums. Come join Lucid Commerce, which is a start up in Pioneer Square created by advertising industry veterans from Aquantive, Microsoft, Overture, Specific Media, and SpotRunner. At Lucid, we are building the first cross-medium robotic advertising agency using deep advertiser integration and **MACHINE LEARNING**. We are already running media in television and very soon online. We are looking for a self-starter that is passionate developing systems that have low latency/high scale requirements and use **MACHINE LEARNING** techniques to optimize. Work on building the display banner ad server where we need to determine the market potential of a request in less than 30ms given a set of inbound attributes about the audience. Or work on the television system which is running daily to determine which commercial spots are predicted to yield the highest amount of revenue given a feature set of hundreds of attributes about the programming, the buying audience's demographics, other similar product's buyer attributes, etc. This person must be able to work with the data mining team in terms of contributing to model design and take it from design in R to implementation in C#/C++. This person must also be able to learn and build statistical unit testing techniques, operational model maintenance, etc.

The ideal skills are: Understanding of HTTP request (headers, cookie setting, logging, etc); OOP concepts/methodology; **MACHINE LEARNING** or statistics background; C # or Java C++; Experience with caching systems (Ehcache, Memcached, Terracotta, etc); Working knowledge of SQL; Working knowledge of multi-variate testing techniques; Advanced degree in CS, statistics, or other strong quantitative degree a plus.

In terms of campus hiring this is a fantastic opportunity to learn an industry and technology from a highly motivated team of experts and potential to join an early stage startup set to grow. There are also limited opportunities for academic publishing.

Location: Seattle

Compensation: Market Rate

Principals only. Recruiters, please don't contact this job poster. Please, no phone calls about this job! Please do not contact job poster about other services, products or commercial interests.

PostingID: 1543312019 Copyright 2010 craigslist, inc.

1.8 JobID

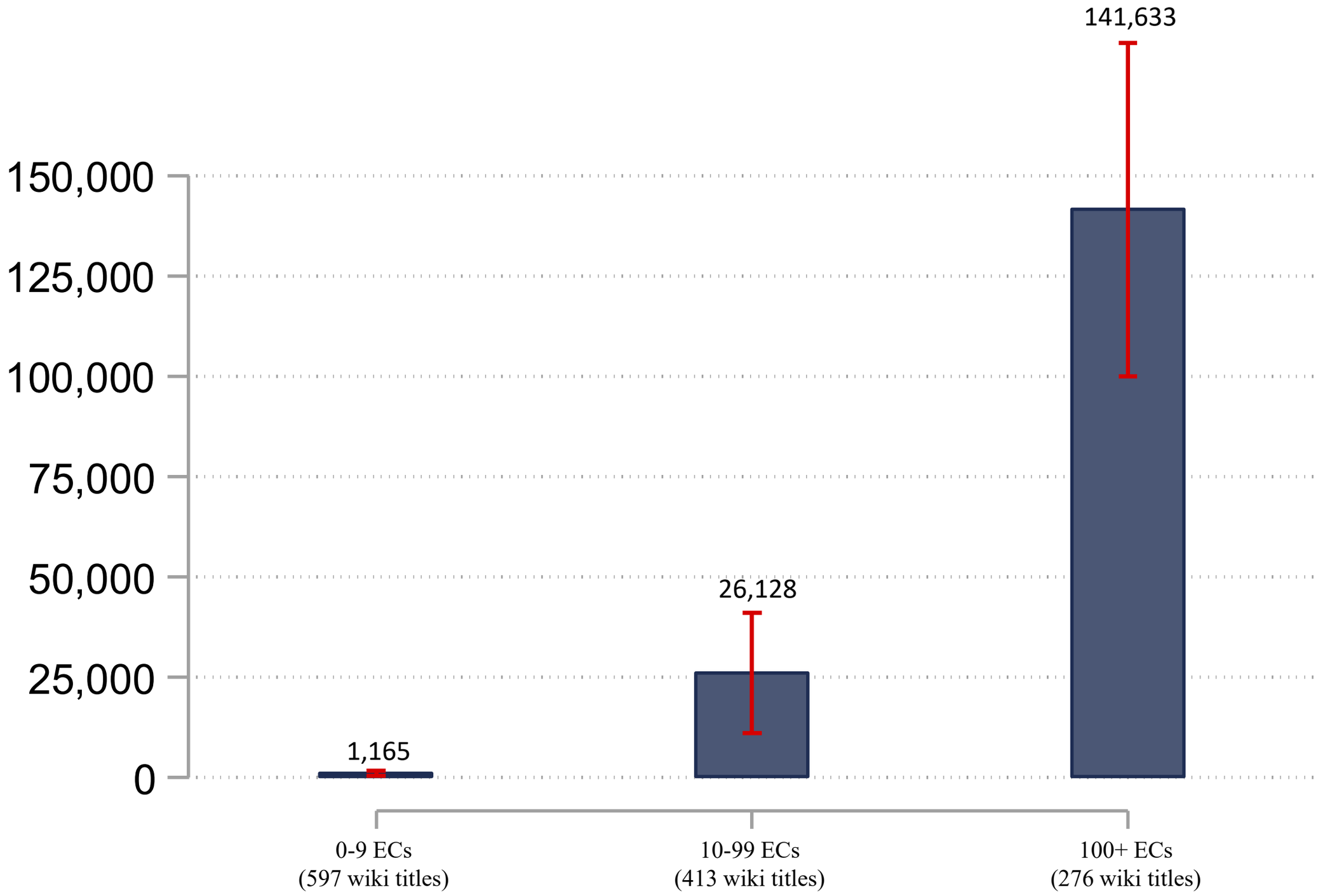
311172646

1.9 JobURL

http://seattle.craigslist.org/see/sof/1543312019.html

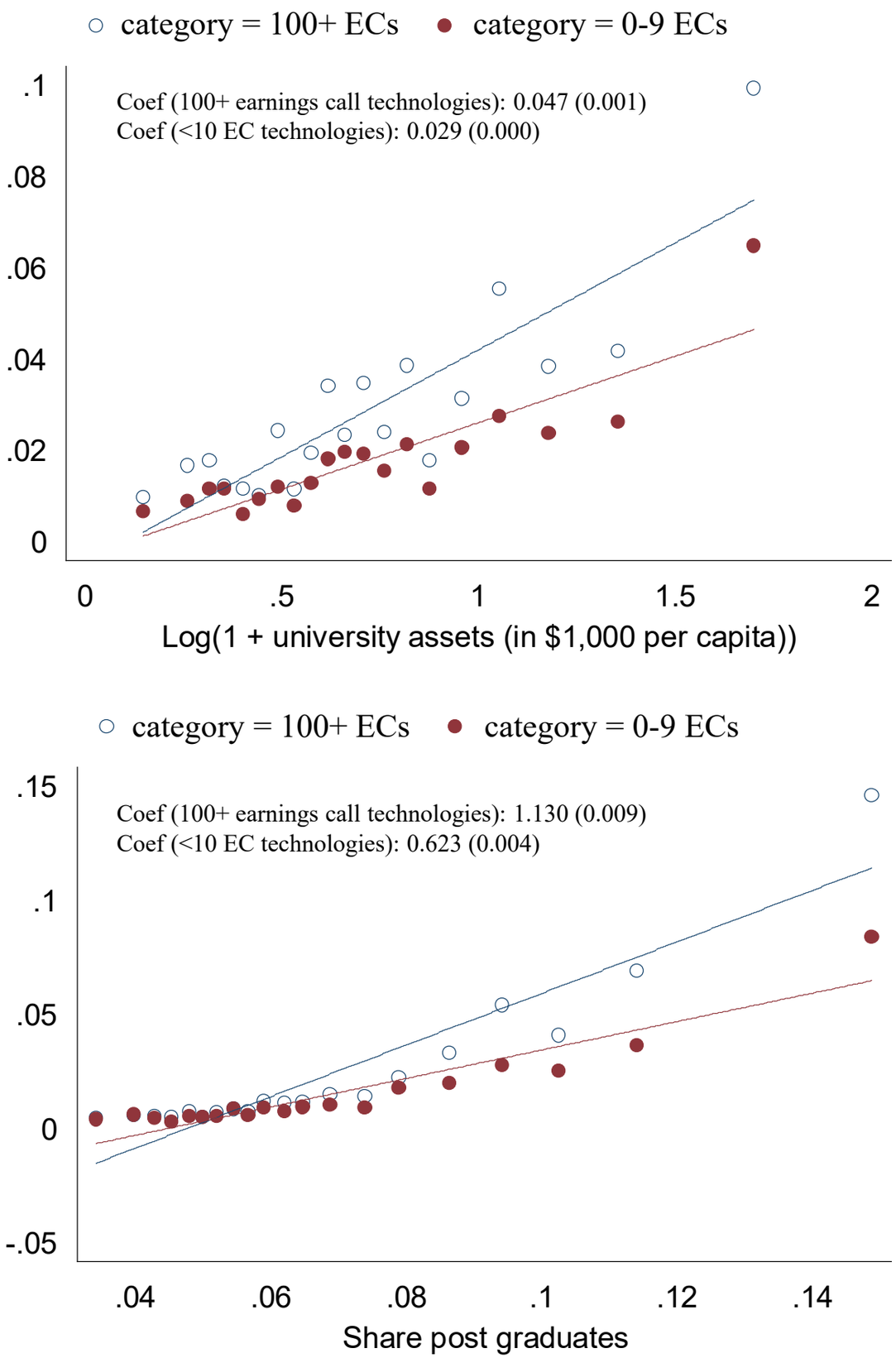
Notes: The figure shows an example of a job posting in Burning Glass, with associated data. This posting mentions one of our technologies ("machine learning").

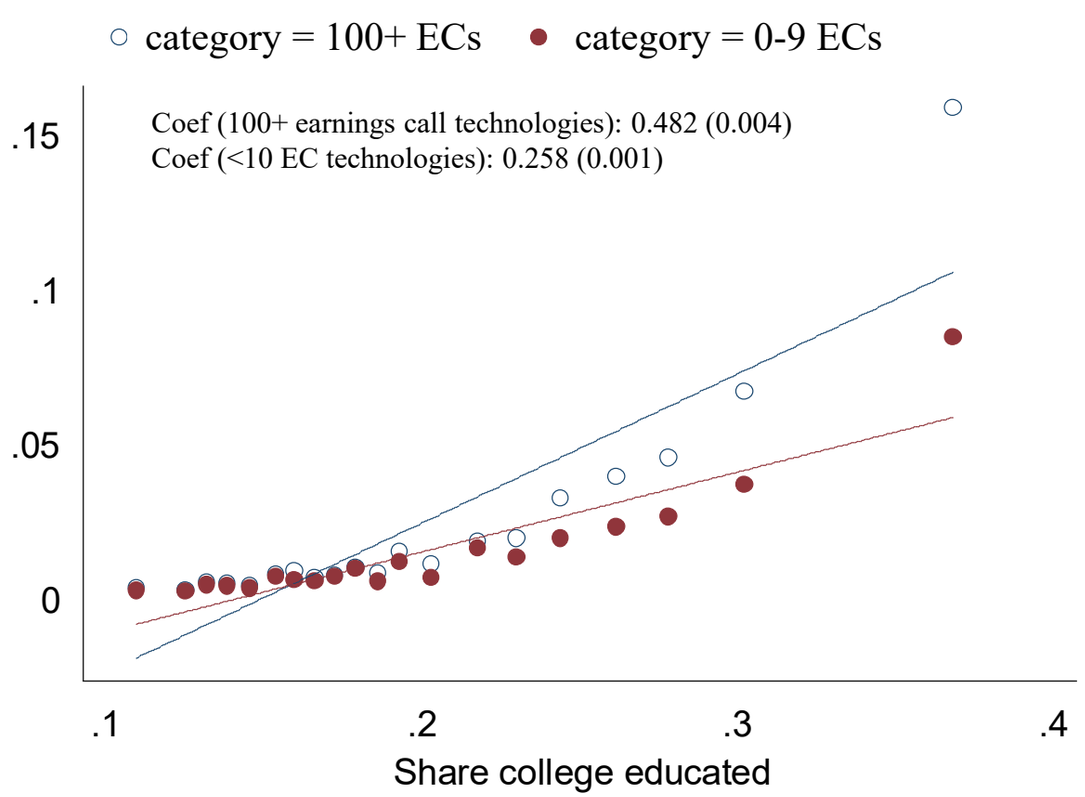
Appendix Figure 4 – Average number of job postings by earnings call mentions associated with the technology



Notes: The figure shows the average number of job postings associated with technologies associated with at least one bigram that appears in at least 100 earnings calls, those technologies associated with at least one bigram that appears in at least 10 and less than 100 earnings calls, and the remainder. Whiskers show 95%-confidence interval. A t-test of difference between the average number of postings results in a F-stat of 57.04 (p-value of less than 0.001).

Appendix Figure 5 – Early patenting vs. local skill composition





Notes: The figures show binned scatter plots of patents in the ten years prior to the emergence date per 1,000 people in a CBSA over measures of skill and university presence in the CBSA. The coefficients and robust standard errors displayed are from the associated regressions with standard errors clustered by CBSA. The figures are shown separately for those with 100 and over and less than 10 earnings call mentions. In all panels, the coefficient for patents are (statistically) significantly higher for those associated with 100 or more earnings calls than the coefficient for the other patents at 1% level.

Appendix Figure 6 – Examples of use and research, develop, and production (RDP) in job postings

Use: Sales Representative

We are in search of an outgoing driven and reliable individual who is looking for a part time or full time opportunity to become a brand rep for a regulated product in convenience store locations in your area you will be a key asset to the program…

Responsibilities and requirements:

- work on product displays pull product out of back stock and merchandise replenish displays as needed.
- use third channel technology on a smart device to collect crucial data engage with consumers and provide sales support/brand education to retail associates
- *reliable transportation a **smart phone** with internet access.*

Use: Clinical Nurse

Position is responsible for creating and maintaining positive customer relations for hospice and, as appropriate, home health and other-in-home services programs throughout the assigned hospital/clinic.

Essential Duties:

- …
- Documents in patient hospital medical record pertinent information related to discharge from hospital to home or facility/agency setting.
- *Appropriately documents activities in the Providence **electronic medical** system, tracks referrals received by nursing unit and accepted by each Providence agency.*

Use: Ultrasound Technologist

The Ultrasound Technologist II under general direction, operates ultrasound equipment to obtain high quality ultrasound examinations of various body parts as ordered by the Physician. Prepares patients, processes images and assists Physicians as needed. Evening Shift and rotate call. Qualifications RDMS required. ARRT a plus. 3-5 years of experience preferred. *Successful completion of a formal (Accredited) educational /training program for Radiology and/or **Medical Imaging**.*

RDP: Software Engineer

Job Description: *Analyze requirements, develop the **software architecture**, plan and perform detailed software design activities, specify software test requirements, estimate software size and development effort and plan for future product releases is required.*

RDP: Scientist

[The company] has an opening in R&D for an individual experienced in radiological imaging to lead the development of computed tomography (CT) systems used in pre-clinical imaging research. We are emphasizing a multi-modality imaging approach, combining CT with optical molecular imaging. The successful candidate will guide the development of both hardware and software platforms, working closely with a multi-disciplinary team of engineers, software developers, and biologists…

*With a keen focus on clinically-relevant experimentation, [this company's] portfolio of offerings includes state-of-the-art microfluidics, lab automation & **liquid handling**, optical imaging technologies, and discovery & development outsourcing solutions….*

RDP: Modeling & Simulation Software Engineer

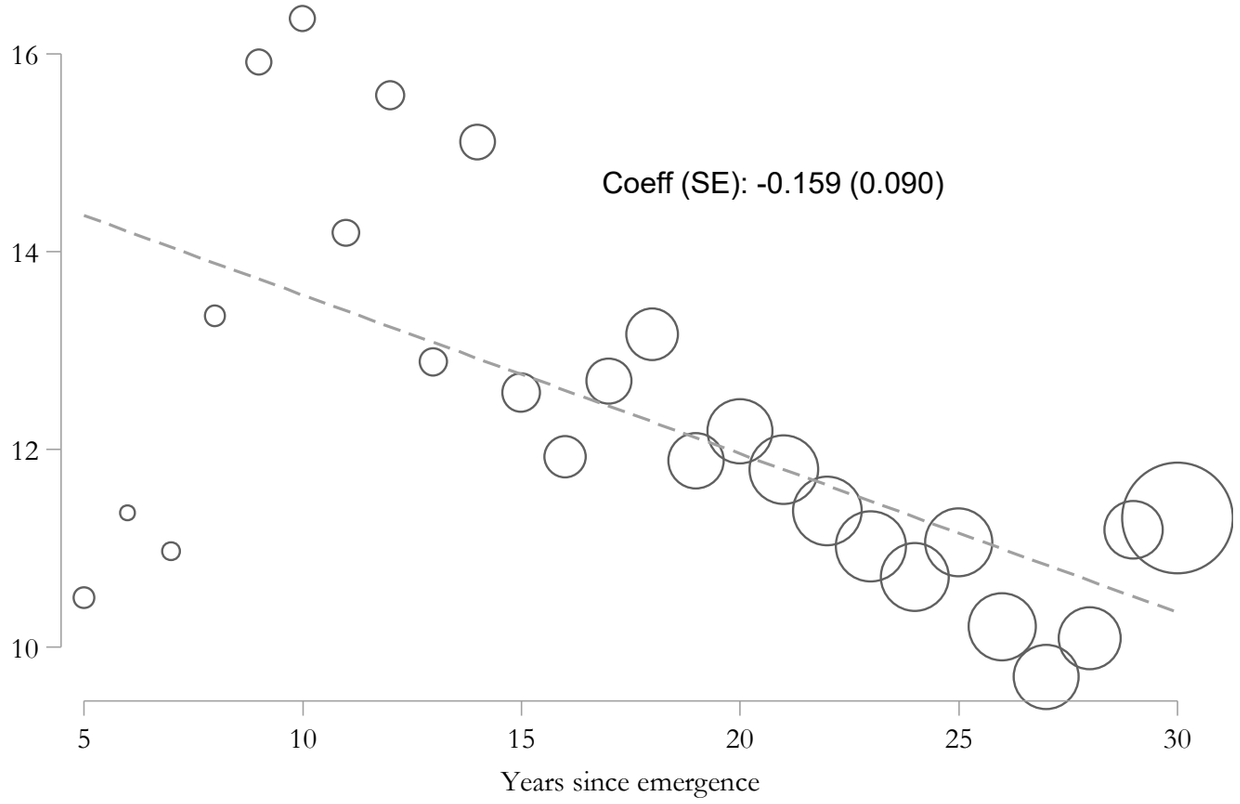
Our client is seeking a motivated and passionate scientist/engineer with the technical vision required to translate advanced modeling and simulation concepts into operational reality. (…) The successful applicant will function as a key technical member of an interdisciplinary team in a dynamic development environment, providing leadership across the full R& D cycle. Position responsibilities will include concept formulation, proposal writing, algorithm design, software implementation and integration, client interaction, and research publication…

Ph.D. in Computer Science or M.S. in Computer Science plus minimum 5 years experience . Hands-on experience with one or more advanced modeling approaches, transforming theory into real world solutions. *Strong software development skills including systems integration, network programming, AI programming, and **virtual environments**.*

Notes: The panels above show examples of job postings mentioning our technology bigrams categorized into use, and research, development, and production.

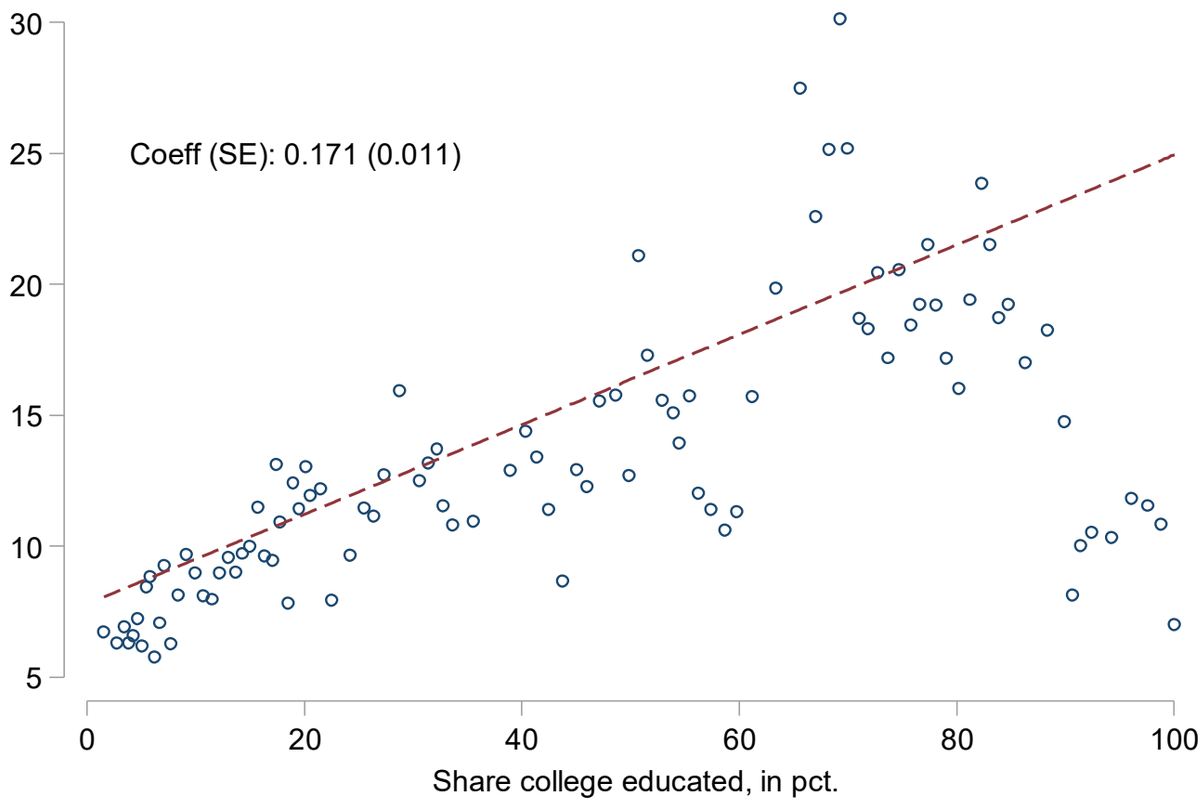
Appendix Figure 7 – Technology job postings with RDP synonyms, by year since emergence and college education

Panel A: Pct. of technology RDP job postings by year since emergence



Notes: For a panel of technology bigram x year observations, the figure in Panel A shows a binned scatter plot of the share of Research, Development, and Production (RDP) job postings relative to the years since emergence. Each bin represents the average share of job postings categorized as RDP jobs across technology bigrams for a given year since emergence. The sample pools across technologies, where each observation in the sample denotes a technology bigram x year observation. Only observations at the year of emergence and after are included. The size of circles is proportional to the square root of the number of technology bigram job postings, capped at 100. Sample includes bigrams which appear in at least 100 earnings calls.

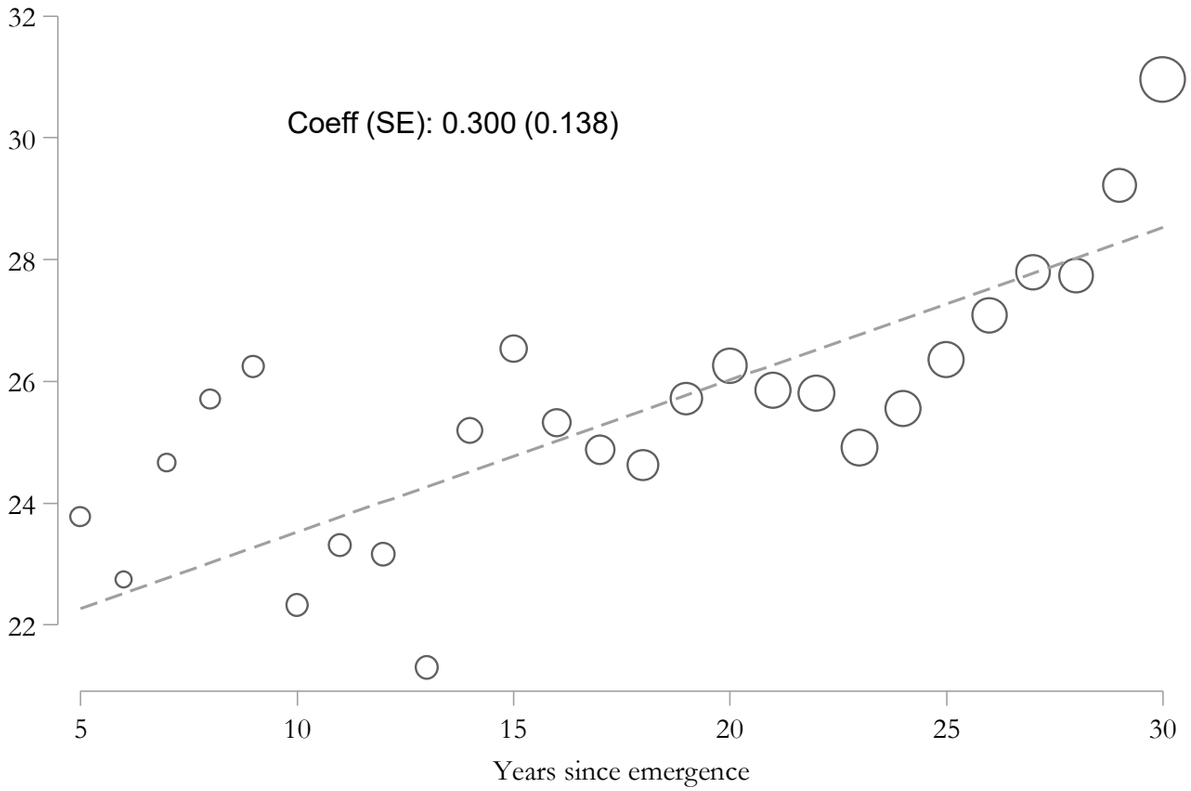
Panel B: Pct. of technology RDP jobs by college education



Notes: The figure in Panel B plots a binned scatter plot of share of Research, Development, and Production (RDP) postings against the share of college-educated postings. Each observation in the dataset is at the technology bigram x year level.

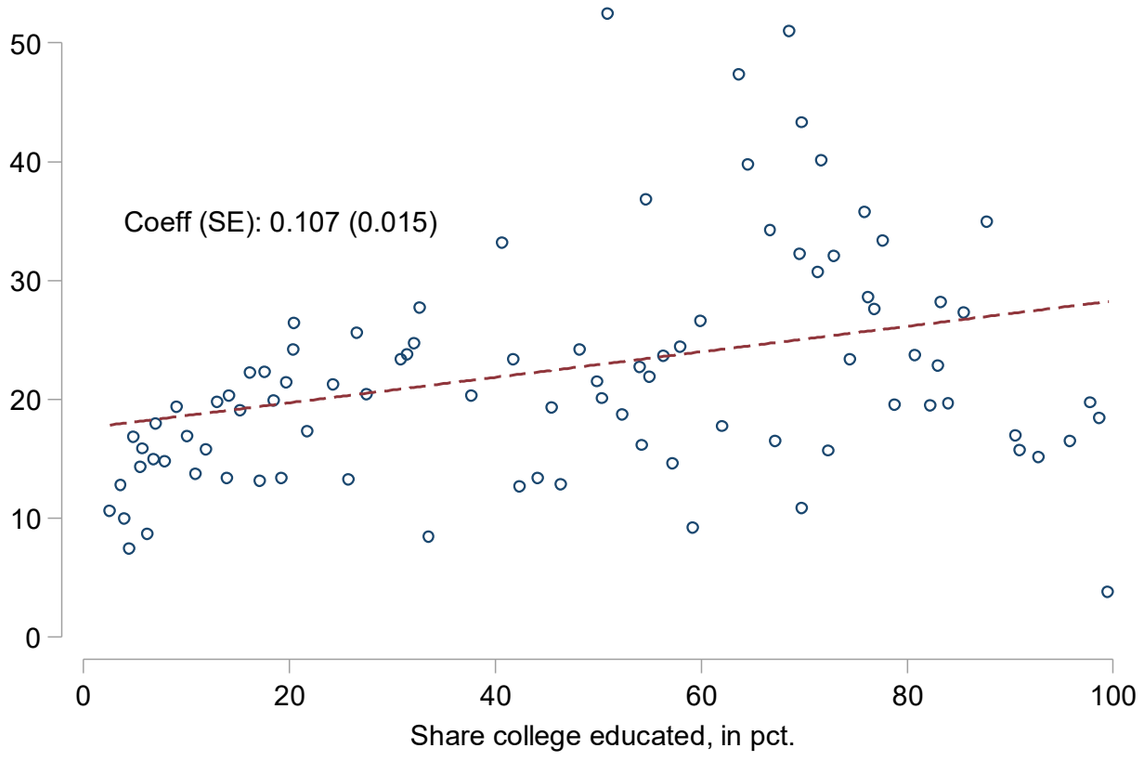
Appendix Figure 8 – Technology job postings with training synonyms, by year since emergence and college education

Panel A: Pct. of technology training job postings by year since emergence



Notes: For a panel of technology bigram x year observations, the figure in Panel A shows a binned scatter plot of the share of training postings against the years since emergence. Each bin represents the average share of job postings categorized into `training' across technology bigrams for a given year since emergence. The sample pools across technology bigrams, where each observation in the sample denotes a technology bigram x year observation. Only observations at the year of emergence and after are included. The size of circles is proportional to the square root of the number of technology bigram job postings, capped at 100. Sample includes bigrams which appear in at least 100 earnings calls.

Panel B: Pct. of technology training jobs by college education

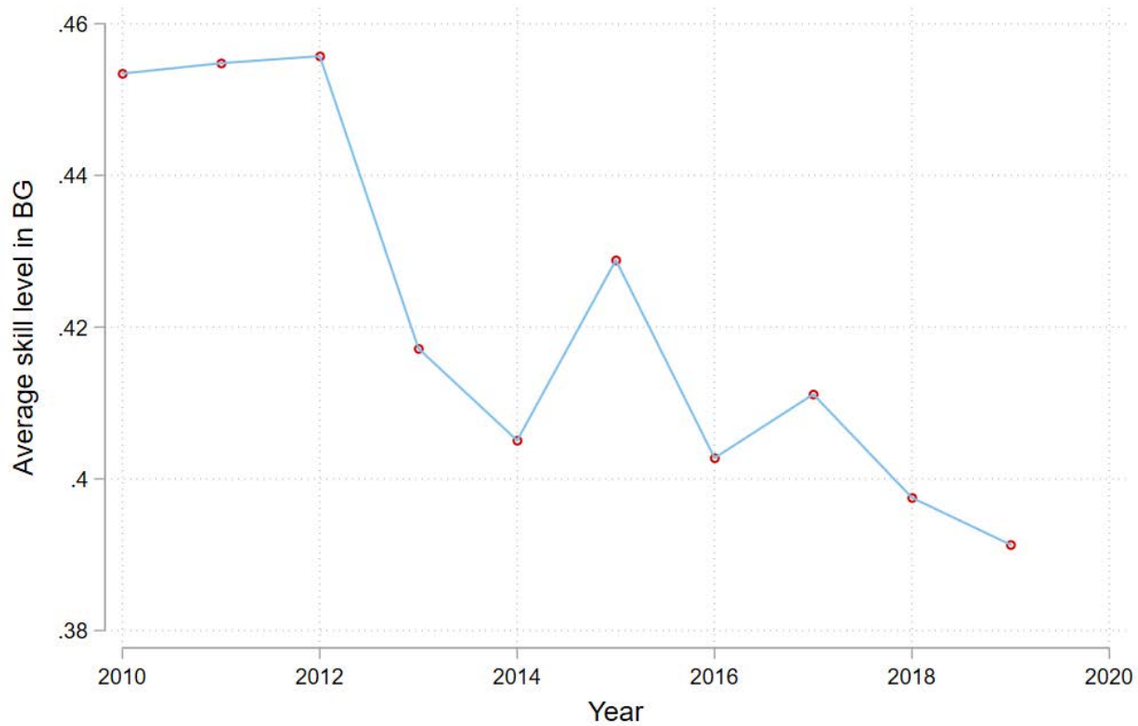


Notes: The figure in Panel B plots a binned scatter plot of share of training postings against the share of college-educated postings. Each observation in the dataset is at the technology bigram x year level.

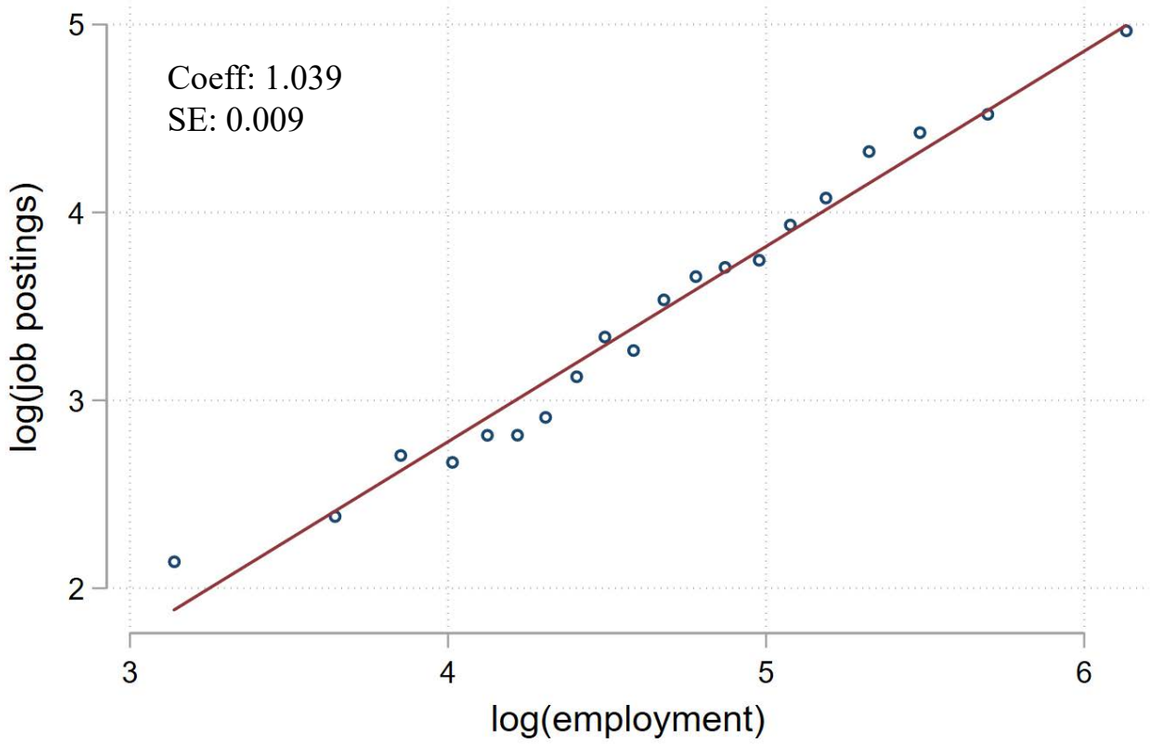
Appendix Figure 9 – Total number and composition of Burning Glass job postings, over time

Panel A: Number of job postings in BG vs JOLTS

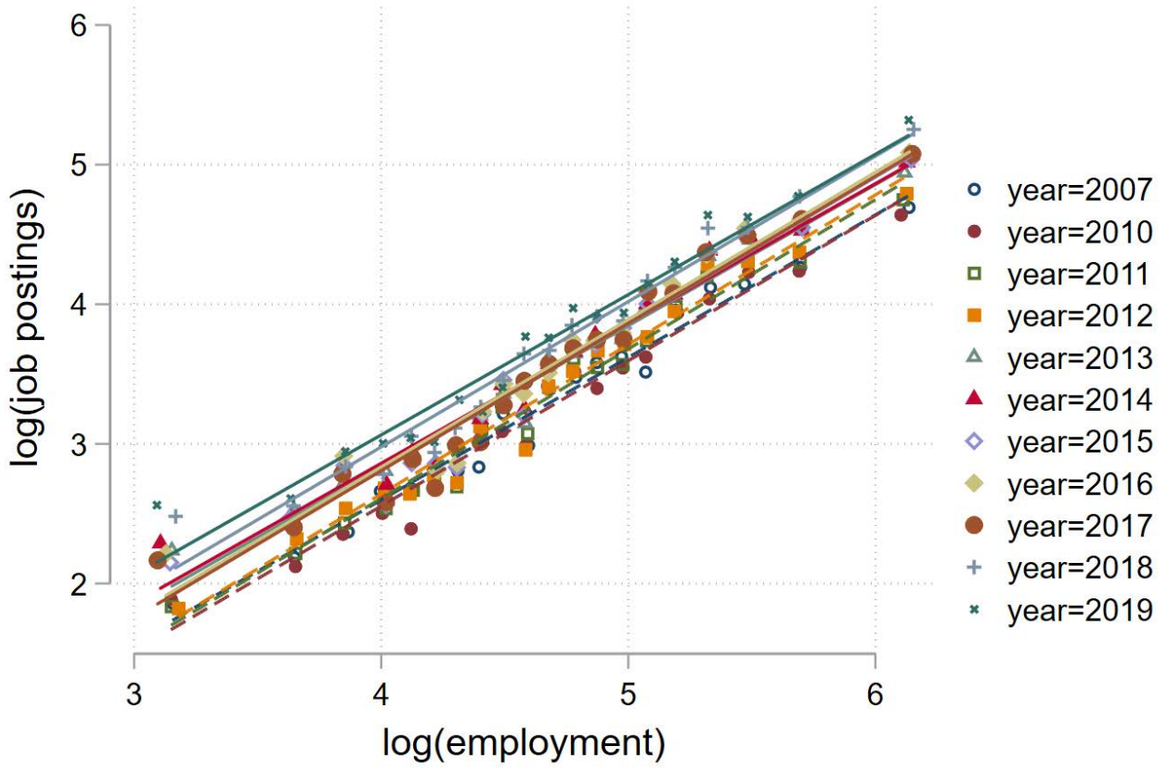
Panel B: Share of job postings requiring a college-education



Panel C: Number of job postings vs employment in OES

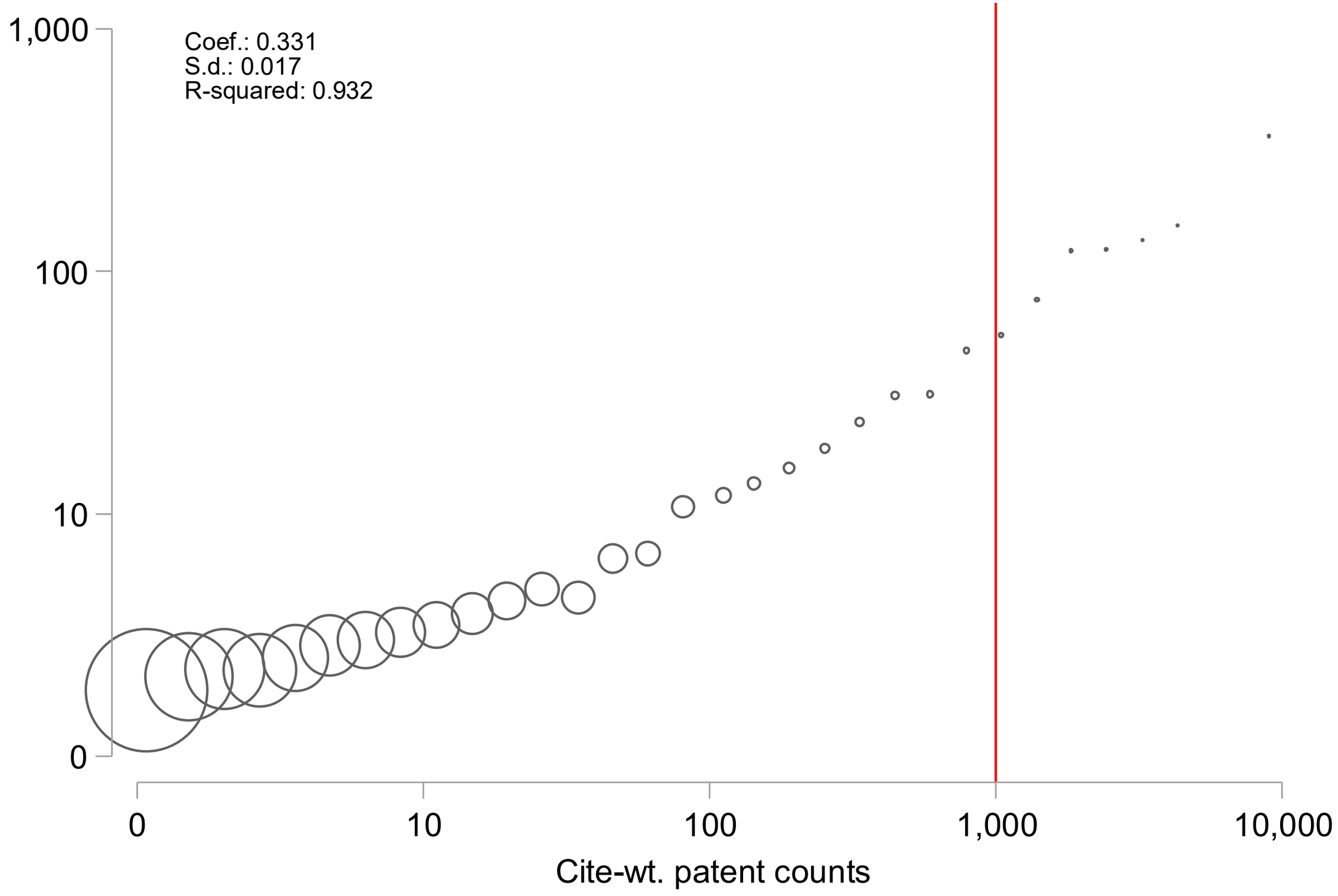


Panel D: Number of job postings vs employment in OES (by year)



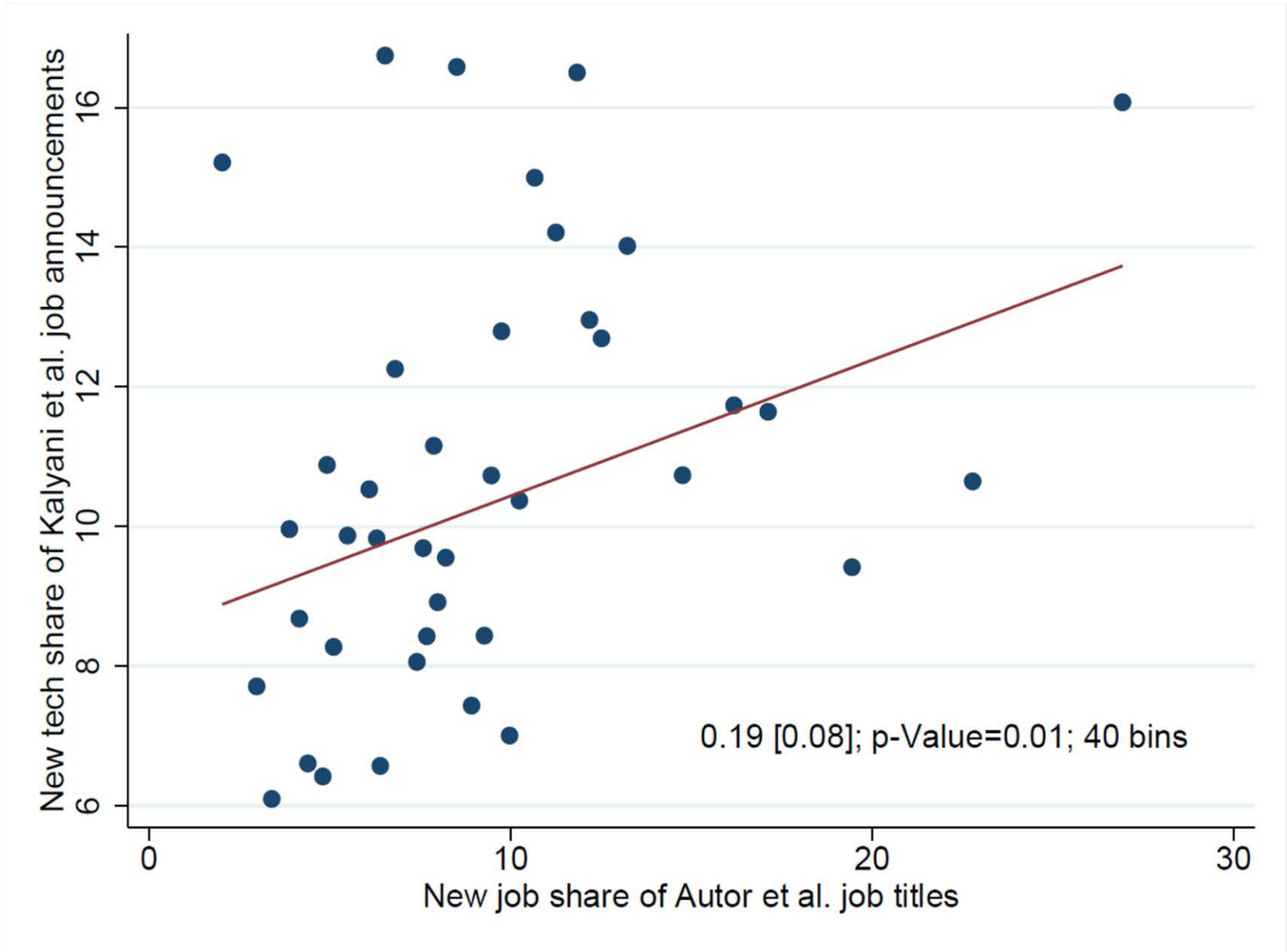
Notes: The figure presents descriptions of the changing composition of Burning Glass (BG) job postings over time. Panel A shows the overall number of job postings in BG (left y-axis) against the overall number of job postings in JOLTS (right y-axis). Panel B shows the (approximate) share of job postings in BG requiring a college education by year. Panel C plots a bin-scatter of number for job postings in BG against the number of employed people in the Occupational Employment and Wage Statistics (OEWS) program by 6-digit occupation (SOC) and year. Panel D plots the same picture as in Panel A with a bin scatter by year.

Appendix Figure 10 – Number of normalized citation-weighted patents and job postings associated with novel bigrams



Notes: The figure plots a binned scatterplot of the number of job postings associated with technologies against the number of normalized citation-weighted patents that mention a novel technology. In preparing the figure, we include all technological bigrams with more than 100 citation-weighted patents, as well as a random sample of 1,000,000 novel bigrams with cite-weighted counts between 0 and 99. The size of circles is proportional to the number of bigrams in any given bin. (Unlike in the other figures and regressions, here we do not cap the weights to highlight the concentration of bigrams on the low end of these measures.) The figure presents a binned scatter plot along with the coefficient, standard deviation, and R-squared from a regression of the log of job postings on the log of cite-weighted patent counts corresponding to each technology.

Appendix Figure 11 – Share of technology jobs and share of new job titles



Notes: The figure plots a binned scatter plot of the relationship between the share of new occupational titles, as measured by Autor et al. (2023) for the years between 1980 and 2018, and the share of job announcements between 2010 and 2019 that are associated with new technologies. Both series are weighted by number of job titles identified by Autor et al. between 1980 and 2018 and are winsorized at the 99th percentile. Observations are at the level of occupational classes in the *1990 occupation code* ("consistent occupational code") classification scheme.