Cathedrals Opening Bazaars*

Placing open-source software in the patent technology space

Xiaofeng Gong

Draft — October 24, 2026

Abstract

When firms make technology public through open-source code, do they reveal the technologies they protect with patents, or a separate set? I link Compustat firms to their USPTO patents and to their GitHub accounts, and a large language model places their open-source projects in the patent classification. In the average firm-year, 35% of a firm's open-source code lies in technologies it does not patent, against 15% under a randomization benchmark built from the firm's own items. The excess of the distance between its open-source code and its patents over this benchmark rises with the count of open-source projects about as much as in simulated firms that keep one technology mix at every size. Outside its patented technologies, a firm's open-source code is less co-located with its nearest technology neighbors' patents than under the benchmark, consistent with the stated prediction that firms reveal less where rivals can use more of what is disclosed. Within a firm, however, new participants make up a larger share of projects inside its patented technologies than of its other projects, unlike the prediction that communities participate less near patented inventions.

1 Introduction

Large firms that file patents now also publish open-source code, which anyone can read, use and build on. Projects run by companies account for a large share of public open-source code (Spinellis et al., 2020). Like a patent, open-source code reveals technology to others (Horstmann et al., 1985; Henkel, 2006). A firm that does both draws a line between the knowledge it shares and the knowledge it protects. Patents are one means of protection among several, and no industry relies on patents alone (Cohen et al., 2000, §3.1, p. 9). Where a firm draws the line between sharing and protecting is a question in the economics of disclosure and appropriability (Teece, 1986).

Existing data do not show where that line runs. Studies of corporate open-source code mostly measure how much code firms open (Lin and Rai, 2024; Cao et al., 2026). Studies of firm technology describe a firm's *technological position* by the distribution of its patents across patent classes (Jaffe, 1986, p. 985), though patents record only part of what a firm knows (Jaffe, 1989, p. 89). Open-source

*The image of the cathedral and the bazaar is from Raymond (2001, p. 21).

code carries no patent class, so it cannot be compared directly with a firm's patents. This paper asks where the line runs. When firms make technology public through open-source code, do they reveal the technologies they protect with patents, or a separate set?

To answer this question, I build a data set that places firms' open-source code and patents in the same classification. Large language models make it feasible to read and match records at scale and at low cost. With their help, I build two crosswalks that link 13,651 Compustat firm records to their USPTO patents granted in 2015–2024 and 5,403 to accounts on GitHub, a public site for software (section 3.2). In two stages, a large language model (GLM-5.3-Flash) reads each open-source project and assigns it classes from the Cooperative Patent Classification (CPC), the scheme patent examiners use (section 3.3). The classes are CPC *main groups*, such as G06F 8/00 (software engineering). When a second language model applies the same two stages to patents, its labels share at least one main group with the examiners' labels on 89% of the patents, against 21% when its labels are assigned at random within each CPC section (section 3.5). I compare a firm's open-source code with its patents using the Bar–Leiponen distance (Bar and Leiponen, 2012, p. 457), the share of the firm's open-source code that would have to move to other main groups to match its patents. I measure the distance for each *firm-year*, using the firm's patents and open-source code over the five years ending in that year (section 4). A distance between two small portfolios is positive even when their technologies are assigned at random. I therefore compare it with the distance under a *randomization benchmark*, which reassigns at random the technologies of the firm's own patents and open-source projects (section 4.2). I call the observed distance minus the benchmark distance the *excess*.

The comparison sample holds 595 large firms that both patent and actively publish open-source code (section 3.6). Their open-source code lies mainly in software engineering and computing, while their patents also cover communications and the physical sciences (section 3.4). In the average firm-year, with each project weighted by its events, 35.1% of a firm's open-source code lies in technologies it does not patent, against 15.1% under the randomization benchmark (section 4.3). Firms' open-source code thus lies in technologies they do not patent more than twice as often as the benchmark places it there. Under event weights, the excess averages 0.2336 (95 percent interval 0.2228 to 0.2443). It is positive in 97% of firm-years, and the excess averaged over a firm's years is positive for 97% of firms (section 4.3). The excess also counts differences in shares inside the main groups the firm patents. With each firm-year counted once, the main groups where the firm holds no patent carry 86% of the mean excess under event weights (97% with the labels of a different model, Appendix C) and 63% under equal weights, which count each project once. When each firm-year is weighted by its open-source events, so that the largest programs count most, their share falls to 6%.

In regressions on the size of a firm's open-source program, the excess is the outcome of interest (section 5.2). Firms with more open-source code have it nearer their patents, but the benchmark distance at each firm's own counts falls faster. The fall in the observed distance thus combines the small-unit bias, which raises the distance between small portfolios and fades as the counts grow, with any change in what the firm opens. Within a firm, randomization inference rejects a fall at the

benchmark's pace (two-sided $p = 0.0020$). A doubling of a firm's open-source projects therefore goes with a rise of 0.036 in the excess. Simulated firms that keep one technology mix of their own at every size show a similar rise of 0.0324 (Monte Carlo $p = 0.169$ for the difference, section 5.2.3). When firms are compared at equal numbers of items, the excess does not rise with the count within a firm, and across firms it falls slightly (section 5.3). The data are thus consistent with each firm keeping one technology mix of its own at every size of its program.

Technologies differ in how often firms open them. Technology norms, the rates at which other firms open each technology, reproduce 49% of the excess (section 4.4). The same rates within the firm's industry add a further 0.7%, with an interval that includes zero, and the remainder is 50%. Firms also differ from one another in how much of their open-source code lies outside their patented technologies, more than the randomization benchmark produces. This dispersion across firms is 0.0937, against 0.0621 under the benchmark, an excess of 0.0316 (plus-one Holm-adjusted $p = 0.0199$), and it lies above its band however patents are dated (Appendix B.6). The dispersion across technologies also lies above its own band, at 0.1261 against 0.1135, an excess of 0.0126. Against the randomization benchmark, firms thus differ from one another beyond it by more than technologies do, and where a firm opens is more a property of the firm than of the technology. Section 5.3 describes how the excess varies with size, seniority and industry.

These findings bear on what firms reveal. Firms reveal selectively (Henkel, 2006, p. 953), and Alexy et al. (2013, p. 272) define selective revealing by knowledge that the firm could have kept proprietary. Henkel et al. (2013, p. 68) address where revealed knowledge sits. When the co-creators of value are distributed, numerous and anonymous (their Table 1, p. 70), Henkel et al. (2013) recommend relatively unprotected open modules beside highly protected closed ones (p. 72). West (2003, p. 1279) describes platform firms that give up control of commodity layers and keep control of the layers that presumably offer more room to differentiate. These works recommend or describe an arrangement, not a prediction for a distance between two portfolios. This paper maps modules and layers onto CPC main groups and reads the condition on co-creators with a proxy, a repository's outside participant-years. These count, year by year, the people active on the repository who were not among its first contributors and had no earlier event on the firm's accounts. Repositories with more outside participant-years keep a slightly smaller share of their code outside the firm's patented technologies than the randomization benchmark gives, not a larger one. Across a firm's repositories in a year, the coefficient on the log of one plus the outside participant-years differs from its mean over the benchmark's draws by -0.004 (standard error 0.0011, Monte Carlo $p = 0.01$, section 5.4). This placement is not consistent with the separation that Henkel et al. (2013) recommend, and the proxy also counts new hires.

One finding matches a stated prediction specific to the question. Outside its patented technologies, a firm's open-source code is less co-located with the patent technologies of its ten nearest technology neighbors than under the randomization benchmark (Appendix E.8). This pattern is consistent with the stated prediction that openness falls as spillovers to rivals rise (Llanes, 2019, Prop. 5, p. 288) and that greater competition tends to make free revealing less likely (Harhoff et al., 2003,

p. 1762). Under the same kind of mapping, technological proximity stands in here for what rivals could use. Bloom et al. (2013, p. 1348) separate technological proximity from product-market rivalry. Among the ten nearest technology neighbors, I compare those that the text-based network industry classification (TNIC-3) of Hoberg and Phillips (2016) lists as the firm's product-market competitors with others matched to them on technological proximity, on the 975 firm-years of 291 firms whose neighbors include both kinds (Appendix E.8). At comparable technological proximity, the pattern is about as strong for neighbors that are not product-market competitors as for those that are, so the data do not single out product-market rivalry over technological resemblance. One stated prediction does not match. Within a firm, new participants make up a larger share in its projects inside its patented technologies than in its other projects, in line with the repositories that draw outside participants. That differs from the prediction, which Lin and Rai (2024, p. 662) summarize from prior research, that projects near patented inventions draw less community participation. The data leave the predictions of defensive publication and of commoditizing a complement undecided (Appendix E.8). Several reasons for opening predict the same signs, so these comparisons cannot tell the reasons apart.

Contributions. The paper makes three contributions. The method places open-source code in the patent classification. A language model reads each project and assigns CPC main groups in two stages, and I check its labels against the European Patent Office's text categorizer. A second model runs the same two stages on patents, and I check its labels against patent examiners' labels. One study assigns research disclosures to CPC classes and compares them with the firm's own patents by relevance (section 2), while this method compares the technological composition of a firm's code with that of its patents.

The data link Compustat firm records to their patents and to their open-source code, and each link records its supporting evidence. I compare the patent links with DISCERN and KPSS (section 3.2). The low cost of running language models allows the search for GitHub accounts to cover 75,293 Compustat firm records.

The measure is a technological distance within a firm, between its own portfolios. Studies of technological distance compare a firm's patents with other firms' patents to see who learns from whom and who competes with whom. I measure which technologies a firm opens, against those it patents, and compare two whole portfolios within a firm, read against a benchmark from the firm's own items. Won (2026c, §3.2, p. 10) notes that a patent-class profile takes the same value for every repository a firm publishes, so it cannot measure whether a release does what a peer sells. Because the labels describe each project's own technology, they avoid that limit. Section 2 relates these contributions to the closest work.

Roadmap. After a review of related work (section 2), section 3 describes the data and the firm-year panel. Section 4 defines the measures and documents that firms also open technologies they do not patent. Section 5 relates the excess to the size of a firm's open-source program, describes how firms differ, reads two arguments in cited sources (section 5.4), and ends with an extension

comparing the firm with its industry (section 5.5). Section 6 concludes with the next question. Appendix E.8 reads each stated reason for opening against the data, and Appendix G examines whether including a firm's open-source code changes which firms lie nearest to it in technology.

2 Literature Review

This paper adds two elements to the closest work. I place open-source code in the patent classification and compare the distance between a firm's open-source code and patents with the benchmark distance. The measurement draws on eight strands of literature. The first four examine why firms release open-source code and how much they release, what firms reveal and what they protect, how patents describe a firm's technological position, and disclosure through publication, which both channels share. The other four develop the tools used in sections 3 and 4. They cover measures built from text and classification, language models as measurement tools, the distance between portfolios, and comparisons of a measure with random allocation. I discuss the closest papers at the end of the section.[1]

Corporate open-source code. This paper compares the technologies in a firm's released open-source code with those in its own patents. Theory identifies several reasons for firms to release open-source code. These include shaping what customers and users adopt (Lerner and Tirole, 2005; Economides and Katsamakas, 2006; Parker and Van Alstyne, 2018) and attracting community contributions (von Hippel and von Krogh, 2003; Nagle, 2018). Developers of open-source code make their code available to everyone on equal terms, which Harhoff et al. (2003, p. 1753) call free revealing.[2] Among empirical studies, Mei (2025, pp. 20–24) interprets corporate open-source code as a labor contract that attracts volunteer developers, and Gortmaker (2025, pp. 3–5) estimates firms' use of and investment in open-source code in equilibrium without a patent channel. Habibi (2025, §3) places firms in a latent patent-technology space without classifying the released models, and Cao et al. (2026, pp. 18–19, 31) relate firms' patent-based technological profiles to the level of their GitHub activity but do not place any open-source project in CPC. Coleman et al. (2025) provide a framework for how firms allocate innovation among open-source code, patents and secrecy. Studies of value examine GitHub adoption and the associated earnings and announcement returns (Coleman et al., 2025, Tables 3, 4 and 7), and the value of open-source projects from market reactions, by topic and license (Emery et al., 2025, Tables 5, 8 and 9). They also examine release events alongside the firm's prior patents (Yang, 2026, pp. 5, 8–10), the productivity of firms that use free open-source code (Nagle, 2019, p. 1191) and the economy-wide value of open-source code (Hoffmann et al., 2024,

[1] Page numbers refer to the versions in the references.

[2] Developers contribute for delayed rewards that signal their talent (Lerner and Tirole, 2002, p. 214). The same firms commonly develop both open-source code and proprietary software (Lerner and Schankerman, 2010, ch. 1). Company-hosted projects are prevalent across the software industry (Osborne et al., 2025, p. 2). Other companies can contribute code to these projects. Their contributions account for about a fifth of commits in three AI frameworks (Osborne et al., 2025, p. 1).

pp. 19–22).[3]

What firms reveal and what they protect. Teece (1986, p. 287) ties an innovator's ability to capture the profits from an innovation to its appropriability regime. Firms protect knowledge by several means, and no industry relies on patents alone (Cohen et al., 2000, §3.1, pp. 7, 9). Henkel et al. (2014, p. 879) name secrecy and legal exclusion rights as the two means of keeping innovation-related knowledge exclusive, and they note that open innovation can also work without exclusion rights. Firms reveal selectively (Henkel, 2006, p. 953). Alexy et al. (2013, p. 272) define selective revealing by knowledge that the firm could have kept proprietary, and their propositions state conditions under which a firm is more likely to engage in it (Propositions 2b and 4, pp. 278 and 281). Theory relates how much firms reveal to competition and spillovers. Greater competition tends to reduce free revealing (Harhoff et al., 2003, p. 1762). In the model of Llanes (2019, Prop. 5, p. 288), openness falls as spillovers to rivals rise, and in the model of Kamien and Zang (2000, Remark 2, p. 1006), firms offset exogenous spillovers by choosing firm-specific research approaches. Two works describe where revealed knowledge sits. Henkel et al. (2013, p. 68) call a system IP-modular when the boundaries of its parts with different intellectual-property status coincide with the boundaries of its modules. When the co-creators of value are distributed, numerous and anonymous (Table 1, p. 70), they recommend relatively unprotected open modules beside highly protected closed ones (p. 72). They also describe the opposite arrangement, in which one module combines strong and weak protection and so extends strong protection to both (p. 77). West (2003, p. 1279) describes platform firms that give up control of a platform's commodity layers and keep full control of the layers that offer more room to differentiate. None of these works states where a firm's revealing sits relative to the technologies it patents. This paper measures that placement.

Technology from patents. Economists describe a firm's technological position by the distribution of its patents across patent classes (Jaffe, 1986, p. 985), and they compare these positions to measure knowledge spillovers and product-market rivalry (Jaffe, 1989; Benner and Waldfogel, 2008; Bloom et al., 2013). Patents record only part of a firm's knowledge. Jaffe (1989, p. 89) expects positions to understate firms' interest in fields in which little is patented, including computers and semiconductors, and Podolny et al. (1996, p. 672) note that unpatented public inventions leave a patent-based network incomplete.

Disclosure through publication. A patent publishes its invention (Horstmann et al., 1985), and open-source code is published by design. Firms may also publish to place an invention in the public domain before a rival can patent it (Bar, 2006; Parchomovsky, 2000).

[3]Others measure gross margins (Singh et al., 2026, pp. 1, 3), contribution events (Aksoy-Yurdagul, 2025), startups' funding (Conti et al., 2025), the growth of decentralized-finance protocols (Landsman et al., 2026) and open generative AI (Zhao, 2025). They also measure open-source code as investment (Korkmaz et al., 2024, p. 1) and the value of a free web server missing from measured software investment (Greenstein and Nagle, 2014, p. 623). Another study examines a firm that rebuilt a product on another firm's project (Haese and Peukert, 2026, pp. 4173–4174).

Measures from text and classification. I use the patent classification to measure which technologies firms put into open-source code. Researchers use text to relate patents to one another (Arts et al., 2018, Kelly et al., 2021, p. 303), to occupations (Autor et al., 2024, p. 1402) and to job tasks (Webb, 2020). Argente et al. (2025, pp. 9–13, 23) match firms' patents by text to their own product introductions. Measured proximities between firms change when the same patents are represented by their text rather than their classes (Arts et al., 2023). Gambardella and Torrisi (1998, p. 445) compare two portfolios of one firm, its patents and its businesses, and Szczygielski et al. (2025) compare two sectors' patent filings in a shared taxonomy.[4] Petralia (2025) links organizations' contributions to open-source code to their patents, using a database that Petralia (2024) documents, and Petralia and Boschma (2026, §2, pp. 9–10) compare each patent that names a project with the same assignee's contributions to it, one patent at a time, using the similarity of code embeddings.

Language models as measurement tools. Economists build measures from text (Gentzkow et al., 2019a), and deep learning now turns text into structured data at scale (Dell, 2025). Hansen et al. (2023, pp. 1–2) classify job postings by whether they allow remote work, using a language model tuned on human labels and checked against them. Estimates built on such a measure are valid only when it is combined with a small validation sample (Ludwig et al., 2026, p. 283). Bello et al. (2026, pp. 2343–2347) measure the similarity between two economists' research using embeddings of their abstracts and check the measure against shared JEL codes and journals. What the embeddings capture is hard to interpret. Studies evaluate language models' ability to assign patents to CPC (Lee and Hsiang, 2020; Emer et al., 2026). I use a language model to place open-source projects in CPC and check the labels against two references built independently of my labels (section 3.5).

The distance between two portfolios. I use the Bar–Leiponen distance, the min-complement distance of Bar and Leiponen (2012), because its metric properties link the six distances (section 5.5) and because it measures the share of one portfolio that would have to move to match the other. Jaffe (1986, p. 986) measures the proximity of two firms by the angular separation of their patent-class vectors, and Bloom et al. (2013, p. 1350) add a Mahalanobis version, in which patent classes that occur together within firms receive greater proximity. Bar and Leiponen (2012, p. 457) propose the min-complement distance and show that it equals half the L_1 distance between the two share vectors, so it is a metric (p. 458). In segregation research, the same measure is the index of dissimilarity of Duncan and Duncan (1955, p. 211).

[4] Studies have placed other documents in patent classes, among them research papers (Nanba and Takezawa, 2009, pp. 27–28), non-patent literature (An et al., 2026, pp. 962, 972–973) and standards (Baron and Pohlmann, 2018, p. 504). Concordances map patent classes onto industries (Lybbert and Zolas, 2014, p. 530, Goldschlag et al., 2020, Dorner and Harhoff, 2018, p. 768) and trademark classes (Abbasiharofteh et al., 2026, pp. 2–3). Studies have classified open-source projects by industry (Xu and Quispe, 2026), topic (Emery et al., 2025, p. 14, Balla et al., 2026, pp. 1–2) or use case (Conti et al., 2025). Others have grouped open-source projects by the libraries they use (Brea, 2024, pp. 5, 14).

Reading a measure against random allocation. I compare the distance between a firm's open-source code and patents with the benchmark distance computed from its own items. Indices built on small counts take positive values even when allocation is random. Herfindahl-type measures built on count data carry a known bias (Hall, 2005), and dissimilarity indices share the problem (Allen et al., 2015). Proximity between firms computed from few patents is biased and imprecise (Benner and Waldfogel, 2008, NBER Working Paper 13322, p. 2). Researchers compare segregation indices with random allocation at the observed unit sizes (Carrington and Troske, 1997, §3.1, p. 405) and industry concentration with plants thrown at a dartboard (Ellison and Glaeser, 1997, §II). Duranton and Overman (2005, LSE Research Online version, pp. 9–10) simulate the reference for the location of an industry's plants, drawing its sites only from those its sector occupies (p. 27). A permutation of the observed units also gives such a reference (Gentzkow et al., 2019b, p. 1318). Even an index corrected against random allocation can still depend on the size of its units (D'Haultfœuille and Rathelot, 2017, p. 62). In technology, such references measure how related activities are. Researchers test lines of business (Teece et al., 1994, p. 6) and patent fields (Breschi et al., 2003, p. 75) against random assignment, and Alstott et al. (2017, p. 443) compare the proximity between patent classes with what it would be in randomized patent records. Here the randomization benchmark is built within one firm, for two of its portfolios.

The closest literature. Four papers relate most closely to this paper, each to a different part. Ganglmair and Kann (2025a, slides 8, 13, 14 and 18) assign Research Disclosure publications to CPC (sub-)classes and compare a firm's disclosures with its own patents by relevance, but they report neither the technological composition of the two channels nor the distance between them. I place firms' open-source code in the patent classification alongside their patents. Ganglmair and Kann also find that, within firms, research disclosures concern older or discontinued lines of research (Ganglmair and Kann, 2025a, slides 18 and 20). Appendix E.8 reports when open-source code appears in a technology relative to the firm's patents. This comparison describes timing rather than relevance. Lin and Rai (2024, pp. 666–669, Tables 4 and 6) study the *Alice* eligibility shock at the firm level and find that activity in open-source code rose as software patenting fell. I measure which technologies firms open rather than how much. Won (2026a) describes the commoditize-your-complement incentive at the market level. The study splits lagged product-market peers (Hoberg and Phillips, 2016) into substitutes and complements by their CPC-subclass patent overlap and pre-event return co-movement (Won, 2026b, June 2026 version, §6.1, pp. 27–28), and an earlier draft takes the overlap with a CPC profile inferred for the released open-source project (Won, 2026a, §6.1, p. 27). The same appendix measures where a firm's own open-source code sits relative to its complementors' patent technologies. Finally, Valentine et al. (2025, pp. 1723–1726) find that firms publish more science in the patent classes where they lag, and the appendix also measures where a firm's open-source code sits relative to the patents of peers whose patenting grows where the firm falls behind.

3 The data: linking and classifying open-source code

I link Compustat firms to their open-source code on GitHub and to their USPTO patents, then use a language model to assign patent classes to the open-source code. The labeling and the two crosswalks are the paper's first two contributions, with the labeling at their center. I describe the sources in section 3.1, then build the two crosswalks and bound what they miss in section 3.2. I construct the labels in section 3.3, describe their patterns in section 3.4, and check them against two references in section 3.5. Section 3.6 describes the firm-year panel and the analysis samples. Figure 1 traces the data from the raw sources to the regression panel.

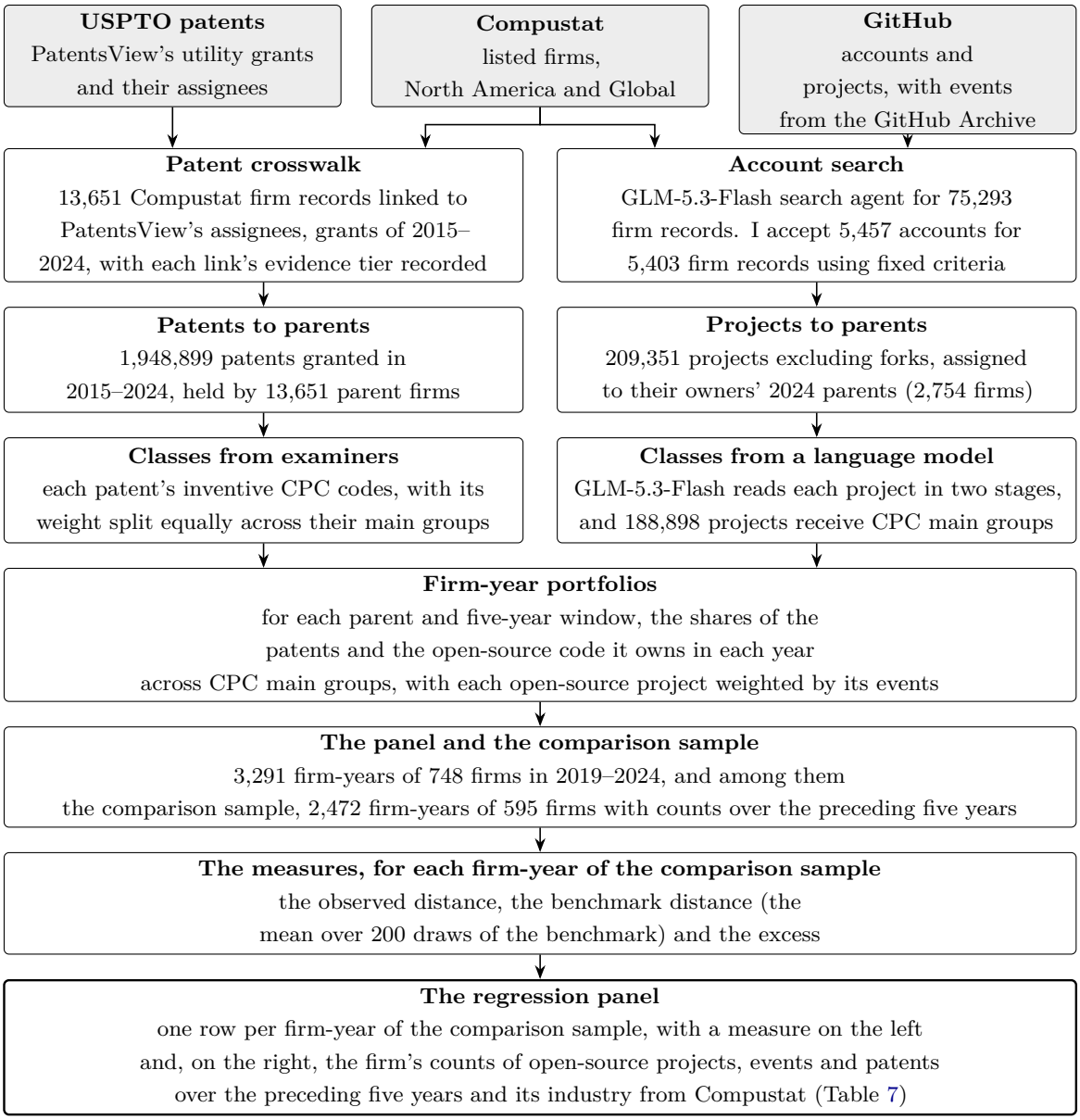


Figure 1: From the raw sources to the regression panel

Notes: Each box shows a step in building the data and the resulting counts. The left column follows the patents, and the right column follows the open-source code. The two crosswalks meet at the Compustat parent firm, where each firm-year's two portfolios are constructed (section 3.2). I assign the labels in section 3.3, construct the panel and the comparison sample in section 3.6, and define the measures in section 4.

3.1 Sources

I combine three data sources at the level of the parent firm. These are open-source projects on GitHub with their activity events from the GitHub Archive, USPTO patents, and Compustat. The parent firm is the company at the top of an ownership tree. I assign its subsidiaries' patents and open-source code to it. Throughout, the firm is the Compustat parent. These are large listed firms whose patents and open-source code are both linked to a parent, and they account for 91.2% of all

GitHub activity that the linked data attribute to a firm. A project's activity consists of its dated events in the GitHub Archive from 2015 onward, regardless of the account that generated them.[5] I weight each project by its events (*event weights*), so projects with more activity receive more weight, and I repeat each result with every project counted once (*equal weights*). I use the firms' utility grants of 2015–2024 recorded by PatentsView.[6] I split each patent's weight equally across the main groups of its inventive codes. Compustat supplies the list of firms and each firm's industry, total assets and R&D intensity. A patent and an open-source project are each an *item* of the firm.

3.2 Crosswalks

Neither Compustat nor the patent records list a firm's GitHub accounts. Spinellis et al. (2020) use committers' email domains to identify GitHub projects developed mainly by companies, but they record no technology classes and no patents. I therefore build the link from firms to their open-source code. I use a search agent to find the GitHub accounts of 75,293 Compustat firm records. I built the agent with Claude Code, using the language model GLM-5.3-Flash and DuckDuckGo's web search. Using fixed criteria, I accept 5,457 accounts for 5,403 of these firm records, and 5,059 of the accounts link to a parent in 2024. The open-source projects under the accepted accounts define a frame of 160,063 projects. I exclude forks, which copy other projects, before linking projects to firms. I identify each project's latest owner in the GitHub Archive and assign the project to that owner's parent in 2024. This procedure assigns 209,351 projects to 2,754 parents. Most projects, 125,461, link to their parent through an owner account that maps to one firm. Another 4,790 link to it through an account that maps to several firm records with exactly the same Compustat name. Another 320 link to it through the owner's aliases in the same year, which all lead to one parent. The last 19 have two defensible parents and count in full for each.

I link patents to parents through a second crosswalk, built separately from the link to open-source code. It links 13,651 Compustat firm records to assignee identifiers in PatentsView, which records USPTO patents and their owners, for utility grants of 2015–2024. Each firm–patent link carries the tier of its strongest supporting evidence. The core tier holds 78.7% of the links. In a further 6.91%, a language model assisted the identification of the firm. Links restored where the published sources thin after 2020, with some added subsidiary links, add 3.23%. Each link requires evidence for the exact firm and assignee identifiers, so similar names alone do not establish a link. For utility grants of 2015–2024, the crosswalk contains 1,948,899 patents in 2,072,295 firm–patent links, held by 13,651 parents.

My crosswalk holds utility grants of 2015–2024, so I compare it with each published source on the years they share. DISCERN (Arora et al., 2021) covers firms headquartered in the United States and their patents granted through 2021, and the extended data of Kogan, Papanikolaou, Seru and Stoffman (2017, KPSS) cover firms in CRSP and their patents granted through 2024. These two

[5] The events are pushes, issues, pull requests, releases, stars, and forks of the project. I use corrected event dates and note when a quantity uses the dates as recorded. The archive's yearly files do not correspond to calendar years. They lack most of January 2018, 2019 and 2020 and include January 2021 twice.

[6] I date each patent by its grant.

published sources link USPTO patents to listed firms. My crosswalk covers 99.7% of DISCERN's patents granted in 2015–2021 and 96.3% of KPSS's patents granted in 2015–2024. Where both link a patent to a Compustat firm, my crosswalk identifies the same firm as DISCERN for 96.6% of the patents granted in 2015–2021 and as KPSS for 97.1% of those granted in 2015–2024. KPSS links each patent to a firm in CRSP, so its Compustat firm is the one that KPSS's own table matches to that CRSP firm. Both sources are among my crosswalk's inputs, so this agreement is partly built in. Among links that do not rely on a published crosswalk, KPSS names the same firm for 92.6% of the patents granted in 2015–2024, and DISCERN for 2.5% of those granted in 2015–2021. DISCERN and my crosswalk use different ownership conventions. DISCERN assigns each patent to its ultimate owner in the grant year, while my crosswalk assigns it to the firm identified by the evidence. A disagreement with DISCERN can therefore reflect this difference or an error in either link. One check rests on a source that never entered my crosswalk, the Global Corporate Patent Dataset (GCPD) of Bena et al. (2017) and Barkemeyer et al. (2026). Where GCPD and my crosswalk both link a patent granted in 2015–2022 to a Compustat firm, GCPD names the same firm for 92.8% of the patents. My crosswalk also covers firms that DISCERN and KPSS omit. It includes 11,173 Compustat firms with patents granted in 2015–2024 that KPSS never links, most of them outside Compustat North America. Of my crosswalk's patents granted in 2015–2024, DISCERN covers 31.8% and KPSS 49.3%. Neither source links firms to open-source code.

Because the search covers 75,293 Compustat firm records, the two crosswalks extend well beyond the firms covered by earlier links. The GitHub crosswalk may still miss firm accounts. In a sample drawn in two strata from patenting firms with no accepted account, I search GitHub's public index using each firm's cleaned name. I find an account with a matching name for 44.4% of them, with a stratified 95 percent interval from 38.8% to 50%. That share is an upper bound, since I do not check the accounts for activity or for a profile that names the firm. The data also omit open-source code that employees release under their own accounts. Gong (2026) documents both crosswalks in full. Both crosswalks are built once. A project counts for its 2024 parent, and the patent crosswalk is undated. The panel then dates ownership. A firm's record ends with its last year in Compustat, whatever the reason, and the panel holds none of its later firm-years. When the record ended in an acquisition or a merger, its later patents and repository-years count for the acquirer that a dated record of a majority acquisition names within a year of the record's end, and otherwise for no firm. Appendix C gives the rule in full, and Appendix C.3 describes the crosswalks' coverage limits.

3.3 Classifying open-source code with a language model

Open-source code carries no patent classification, so I assign patent classes to compare it with a firm's patents. A prompted language model can assign main groups to any project, needs neither word lists nor labeled examples, and runs in about an hour at this scale. Appendix Table C1 compares it with other methods in terms of their inputs, their coverage of varied open-source projects and their cost. I therefore label the projects with a language model and use the EPO's categorizer to check the labels (section 3.5).

A language model reads each open-source project the way an examiner reads an application, and it answers in the patents' own classification. I use GLM-5.3-Flash. It reads the project's name, description, topics, programming language and README. I truncate long READMEs (Appendix C). I prompt the model in two stages. In the first stage, I offer 206 CPC subclasses. These are the most common primary subclasses on patents, supplemented by 7 software subclasses that this ranking would exclude. The model returns at most 3 of them, ranked by relevance. In the second stage, I offer the main groups under the chosen subclasses, and the model returns at most 3, again ranked by relevance. The two stages keep each list short and tie each main group to its subclass. The model sees each project's programming language, so the prompt asks it to classify the project by its function rather than its programming language. Either stage may return no applicable class, leaving the project unlabeled.

I divide the project's events equally across the main groups that the model returns, so the ranking of the labels does not affect their weight. I run the model at temperature zero with low reasoning effort, using the CPC titles from August 2026. Of the 209,351 projects linked to firms, 188,898 receive at least one main group. The other 20,453 receive none and carry no weight in the firm's portfolio. They account for 0.9% of the events in the comparison sample, the firm-years for which the paper compares the two channels (section 3.6). Appendix C.4 follows one project through both stages.

Microsoft's editor for programmers, `microsoft/vscode` on GitHub, illustrates the labels. The model reads its name, description, topics, programming language and README, and it returns two main groups, G06F 8/00 (arrangements for software engineering) and G06F 9/00 (arrangements for program control). Each of the two receives half of the project's events. Given the same project, the EPO's text categorizer returns G06F 8/00.

3.4 What the labels show

Breadth. A labeled open-source project spans fewer main groups than a patent. Among the items of 2015–2024 held by the 748 firms in the firm-year panel (section 3.6), a patent carries 3.01 distinct main groups on average, and a labeled open-source project carries 1.78, or 1.97 under event weights. Breadth affects the distance, as does the number of items (section 4.2). If the two channels recorded the same technology over different sets of main groups, a firm's open-source code and its patents would differ even if what the firm opens did not depend on technology (Appendix A.7). Section 4.4 returns to breadth.

What firms put into open-source code. Firms' open-source code lies mainly in software engineering and computing, while their patents span computing, communications and the physical sciences. CPC section G, physics, which includes computing, holds 88.67% of the firms' open-source code and 39.32% of their patents. Section H, electricity, which includes communications, holds 39.66% of the patents, and sections A to F, the physical sciences, hold 21.02% (Table 1). The open-source code is concentrated. Four main groups hold 60.4% of it, and one main group alone,

software engineering (G06F 8/00), holds 27.61%. The patents' largest main group is information retrieval (G06F 16/00), with 3.82%. The pattern is stable over time. Software engineering's share of open-source code is 28.47% in the window ending in 2019 and 27.52% in the windows ending in 2020–2024. Large firms' open-source code thus sits in a few software main groups, while their patents span many.

Table 1: The distribution of open-source code and patents across patent classes

Technology		Share of open-source code	Share of patents
Panel A. Main groups			
G06F 8/00	Software engineering	27.61%	0.84%
G06F 9/00	Architectures and processors	16.33%	2.6%
G06F 16/00	Databases and retrieval	9.12%	3.82%
G06F 11/00	Error detection and correction	7.35%	1.83%
G06N 20/00	Machine learning	4.68%	0.49%
H04L 67/00	Network protocols	4.01%	1.98%
G06F 3/00	Input and output	1.71%	3.77%
Top four main groups		60.4%	
Panel B. Sections			
G	Physics, including computing	88.67%	39.32%
H	Electricity, including communications	10.73%	39.66%
A–F	Physical sciences and mechanics	0.6%	21.02%
Open-source code outside the patent support		9.62%	

Notes: The technologies are main groups and sections of the Cooperative Patent Classification (CPC). The table pools the five-year windows of the firm-year panel's 748 firms, with each channel pooled over the windows that meet its own condition. The panel retains a five-year window only if it contains at least five open-source projects and at least five patents. The patent floor is the panel's condition of at least five patents in a window. The open-source code column covers every window with at least five open-source projects, including those below the patent floor. The patent column covers every window with at least five patents, including those with fewer than five open-source projects. Each open-source project is weighted by its events, and I split each patent's weight equally across its main groups. The two share columns report the technology's share of the pooled open-source code weight and of the pooled patent weight. I report a technology only when at least three firms have activity in it. Open-source code outside the patent support is the share of the open-source code weight in main groups where its firm holds no patents in the same window, pooled over firms, so the windows with the most open-source code weight count most. It is not the firm-year mean of section 4.3, 35.1%, which counts every firm-year of the comparison sample once. A window below the patent floor counts all of its open-source code as outside. The concentration is stable from the window ending in 2019 to the windows ending in 2020–2024. The top three main groups are the same, and G06F 8/00 holds 28.47% of the open-source code weight in the first window and 27.52% in the later windows.

This concentration is why I measure at the level of the main group (Figure 2). At the level of the CPC subclass, one subclass, G06F, holds 72.49% of the firms' open-source code, and the two largest hold 81.97%. A distance built at that depth would have little room to vary on the open-source side, since most of the open-source code falls in one cell. At the main group, the next level down, the same open-source code spreads over 681 main groups, and it takes 11 of them to reach 80 percent.

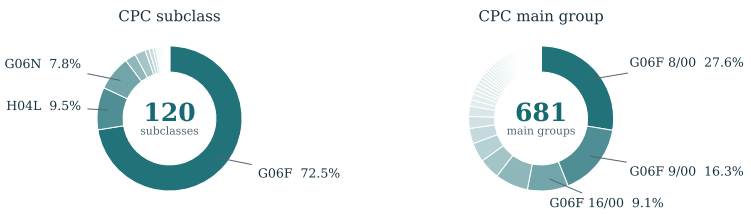


Figure 2: The firms' open-source code at two levels of the patent classification

Notes: Each ring shows the distribution of the firms' pooled open-source code weight across technologies at one level of the Cooperative Patent Classification (CPC), the subclass on the left and the main group on the right. The weight is pooled as in Table 1, over the five-year windows of the 748 firms in the firm-year panel, with each open-source project weighted by its events, using corrected event dates. The number in the center counts the subclasses or main groups that hold any open-source code. Wedges run clockwise from the top, largest first, and the three largest are labeled with their shares.

3.5 Assessing the labels

I check the labels against two references constructed independently of my labels (Table 2). The first is the EPO's text categorizer. My labels share at least one main group with its labels for 76.4% of the open-source projects that both classify, against 38.5% when I assign my labels at random within the reference's CPC section. The second reference is the examiners' classifications. Another language model, DeepSeek-V4-Flash, applies the same two stages to the titles and abstracts of patents, and its labels share at least one main group with the examiners' labels for 89.1% of them, against 20.8% at random. Appendix C reports the first check in full and further checks of the labels.

Table 2: The two label checks

			Share with a common main group	
Reference	Items labeled	Items	Observed	At random
The EPO's text categorizer	Open-source projects, by GLM-5.3-Flash	721	76.4%	38.5%
Patent examiners	Patents, by DeepSeek-V4-Flash	2,774	89.1%	20.8%

Notes: Each row compares labels from the two-stage prompt with a reference built independently of them. In the first row, the labels are this paper's labels for open-source projects, and the reference is the European Patent Office's text categorizer. In the second row, another language model applies the same two stages to the titles and abstracts of patents, and the reference is the examiners' classifications. The observed share is the fraction of items whose labels share at least one main group with the reference. For the share at random, I assign the labels to the items at random within the reference's CPC section. Appendix C reports the first check in full.

I use the labels only in portfolios because the checks measure agreement with other classifications rather than accuracy. Each portfolio is a set of shares built from many open-source projects and compared with the firm's patents at the level of the main group. No result rests on the label of a single project. Label sets also change between runs of the same prompts, while the portfolio shares

stay stable (Appendix C). The results assume that label errors are unrelated to the firm's patent technologies and to the channel. The checks do not test this assumption.

3.6 The firm-year panel and its samples

The unit of analysis is the firm-year, which I construct from the crosswalks and labels. For each firm-year, I collect the firm's patents and open-source code over the five years ending in that year. Open-source code and patents are the firm's two *channels*. The panel contains 3,291 firm-years of 748 firms in 2019–2024. Under the North American Industry Classification System (NAICS), most firm-years in section 5.2 are in the information (38.2%) or manufacturing (32.7%) sector. The *comparison sample* consists of the panel's firm-years that also have counts over the preceding five years. It has 2,472 firm-years of 595 firms and provides the sample for the distance and the randomization benchmark in section 4. Its median firm-year has 41.0 open-source projects with 6,908 events, and 83.5 patents (Appendix Table C3).

Each regression uses the largest eligible subset of the comparison sample. Appendix Table C4 reports the exclusions at each step. Across all firms and windows, open-source code appears in only 723 main groups. I call them the classes that carry open-source code. The regressions in section 5.2 require only the counts and the patent share outside those classes. The regressions in section 5.5 compare the firm with its industry and require an industry classification and the six distances defined there. Table 3 reports summary statistics for the two regression samples, the firm-years in section 5.2 and the firm-years with an industry in section 5.5.

Table 3: The regression samples

	Section 5.2			Section 5.5		
	Mean	SD	Median	Mean	SD	Median
Panel A. Counts over the five years ending the year before						
Open-source projects	174.8	705.3	35.0	176.6	710.0	36.0
Events per open-source project	374.5	1087.3	159.3	375.1	1087.6	160.2
Patents	1026.0	3458.2	74.0	1038.5	3480.8	77.0
Patent share outside classes that carry open-source code	0.172	0.251	0.044	0.170	0.248	0.044
Panel B. The firm, in the year before						
Panel age	11.1	3.5	12.0	11.2	3.4	12.0
Panel C. The firm's own distance, event weights						
Distance	0.8044	0.1650	0.8436	0.8046	0.1642	0.8436
The benchmark distance	0.5715	0.1777	0.5620	0.5714	0.1779	0.5610
The excess	0.2329	0.1359	0.2344	0.2333	0.1357	0.2351
Firm-years		2,447			2,413	
Firms		589			585	

Notes: SD is the standard deviation. Appendix Table C4 describes the samples. This table reports levels of the count variables used in the regressions. The counts cover the five years ending in the year before the distance's window ends. The patent share outside classes that carry open-source code is the share of the firm's patents outside the main groups that carry any firm's open-source code. Panel age is the number of years since the firm entered the panel. The observed distance is the Bar–Leiponen distance between the firm's open-source code and its patents under event weights. The randomization benchmark reallocates technologies among the firm's own items. The benchmark distance is the mean distance over the benchmark's draws. The excess is the observed distance minus the benchmark distance.

4 The analysis data

I examine whether firms publish open-source code in the technologies they patent. This comparison bears on how open-source code relates to a firm's patents. If a firm used open-source code as a direct substitute for its own patents, its open-source code would lie only in technologies the firm patents. In section 4.1, I measure the distance between a firm's open-source code and its patents in each firm-year. Section 4.2 defines the randomization benchmark built from the firm's own items. Section 4.3 compares the observed distance with the benchmark distance, and section 4.4 documents what firms open instead. Section 5 then examines how the comparison varies with the amount of open-source code a firm publishes, across firms and relative to the firm's industry.

4.1 The Bar–Leiponen distance

I use the Bar–Leiponen distance (Bar and Leiponen, 2012, p. 457) to compare a firm's open-source code with its patents. For firm i over the five years ending in year t, let p_{ict} be the share of patents

in main group c and o_{ict} the share of open-source code in that main group. The distance is

$$D_{it} = 1 - \sum_c \min\{p_{ict}, o_{ict}\} = \tfrac{1}{2} \sum_c |p_{ict} - o_{ict}|.$$

D_{it} is the share of the firm's open-source code that would have to move across main groups to match its patent shares. It is zero when the shares match in every main group, and one when the portfolios share no main group. I use this measure to compare firms and relate the distance to firm characteristics. It also measures how unevenly a firm distributes its technologies between open-source code and patents. It is zero when the firm puts the same share of its activity into open-source code in every technology it holds, and it grows as those shares spread apart (Appendix A.3).

4.2 The randomization benchmark

I construct the *randomization benchmark* and the other references used to interpret the observed distances. A distance between two portfolios is positive when they hold few items, even if their items are allocated at random, so its level alone says little about how much they differ (Carrington and Troske, 1997). If what a firm opens did not depend on the technologies of its items, its observed distance would be another draw from the randomization benchmark built from the firm's own items.

The randomization benchmark reallocates the technologies of the firm's items at random among items with the same *start year*, the year an item is first seen. Each patent and open-source project remains in its channel, and each channel retains its number of items and their weights. Each item's set of main groups moves together, and each draw uses each of the firm's sets of main groups once. I compute the distance for each draw as I do for the observed data. The mean distance over the benchmark's 200 draws is the firm's *benchmark distance*. The observed distance minus the benchmark distance is the *excess*, $\tilde{D}_{it}$.

The randomization benchmark serves the same purpose as the dartboard in Ellison and Glaeser (1997). It provides a reference built from the firm's own counts. An analytical expression gives the mean distance when items are allocated independently (Appendix A.8), but it holds only when each item carries one main group with equal weight. The firm's items carry several main groups and weights of their own, and the randomization benchmark allocates a fixed set of them, so its draws depend on one another. I therefore simulate the benchmark distance, as Duranton and Overman (2005, LSE Research Online version, pp. 9–10) simulate their reference for the location of plants. The benchmark distance is the mean distance that would arise if technology played no part in what a firm opens (Appendix A, Lemma 10). For firms simulated under this condition, the mean excess is 0.0008 under equal weights and 0.0009 under event weights. A randomization benchmark could instead draw the firm's technologies from all of its industry's items. That benchmark would hold the industry's mix fixed rather than the firm's, so the excess would also reflect how the firm's technologies differ from its industry's. Section 5.5 asks that question with the industry-pool benchmark.

What the excess measures. The excess measures how much further apart a firm's open-source code and patents lie than the randomization benchmark places them at the firm's own counts. It is positive when the observed distance exceeds the benchmark distance, negative when the observed distance falls below the benchmark distance, and zero on average if what a firm opens does not depend on the technologies of its items. It has the same units as the distance, a share of the firm's open-source code. The firm's *patent support* consists of the main groups in which it holds patents, and the excess splits exactly into open-source code outside the patent support and a mismatch inside it (Appendix A).

Technology norms. Technology norms provide an additional reference for interpreting the excess. Technologies differ in how much of their activity firms publish as open-source code. A firm that holds technologies with different rates would show an excess even if it matched each technology's rate among the other firms. To measure this, I redraw the randomization benchmark so that each item goes into open-source code at its technology's rate among the other firms, the item's *technology norm*. Duranton and Overman (2005, LSE Research Online version, p. 27) build a reference of this kind when they draw an industry's sites only from those its sector occupies. Technology norms play the role of the technology effect in a two-way decomposition, which splits a firm's rate in each technology into a firm effect and a technology effect. Abowd et al. (2002) use a decomposition of this kind to split wages into worker and firm effects. Appendix B.3 reports two further references, the same redraw at the technology's rate among the other firms in the same industry and the fixed-portfolio world, in which each firm keeps its portfolios over the whole period. Table 4 describes each reference, including the two that compare the firm with its industry (section 5.5).

Table 4: The references

Reference	What is held fixed	What is redrawn	Where it is used
The randomization benchmark	The firm's items, their channels, years and weights	Which of the firm's own technologies each item carries, among items with the same start year. 200 draws	Sections 4.3 and 5.2
Technology norms	The benchmark's items, channels, years and weights, and the firm's number of open-source projects	Which items become open-source projects, each at its technology's rate among the other firms	Section 4.4
The industry-pool benchmark	The firm's number of items and weights in each channel	The firm's technologies in each channel, drawn from the industry-year's pooled items, including the firm's own. 499 draws	Section 5.5
The within-industry benchmark	Every item's owner, channel, year and weight	The technologies of the industry-year's items, across channels and owners. 499 draws	Section 5.5

Notes: Each reference holds the firm's counts fixed. It therefore compares a portfolio with a draw of the same size. The industry's two references draw from pools that include the firm's own items. Appendix A.14 gives their exact definitions.

4.3 Not the same technologies

In the average firm-year of the comparison sample, the share of a firm's open-source code that lies in technologies the firm does not patent is more than twice the share under the randomization benchmark. Under event weights, 35.1% of the firm's open-source code lies outside its patent support, against 15.1% under the randomization benchmark. The difference between the two shares is the part of the excess that lies outside the patent support, and the rest of the excess is a mismatch inside it (section 4.2).

In the comparison sample, the excess under event weights averages 0.2336 (firm-clustered 95 percent interval 0.2228 to 0.2443). The interval excludes zero, which is the excess the randomization benchmark gives on average. The share of a firm's open-source code that would have to move to other main groups to match its patent shares exceeds the corresponding share under the randomization benchmark by 23.36%. For the median firm-year, the excess corresponds to 12.9 open-source projects, counting each project once. The excess is positive in 96.5% of firm-years. The observed distance averages 0.8055, compared with a benchmark distance of 0.5719, and is 1.41 times as large. The benchmark distance thus accounts for most of the observed distance. The excess accounts for 29% of the observed distance. Under equal weights, the excess is 0.3304 (0.3195 to 0.3414) and is positive in 99.4% of firm-years. To assess the size of the excess, I compare it with *patent drift*, the change in a firm's patent shares over five years. It is the Bar–Leiponen distance between the firm's patents of 2015–2019 and its patents of 2020–2024, for the 353 firms observed in both 2019 and 2024. When I compare each distance with its own benchmark distance, the excess between a firm's open-source code and its patents is larger than the excess for patent drift by 0.2253 under equal weights (standard error 0.0087) and 0.1266 under event weights (0.0091). Relative to the randomization benchmark, a firm's open-source code lies further from its patents than its own patents drift in five years. The excess persists with one-year and three-year windows, when I exclude patents shared by two firms, when I restrict the sample to projects with an open-source license, and at coarser levels of the CPC. It also persists without the software-engineering main group and with each item assigned to one main group (Appendix B.3 and Table B40).

The observed mean distance exceeds the mean distance in each of the benchmark's 200 draws, whose means lie between 0.564 and 0.579 under event weights (Appendix Figure A2).

In nearly every firm, open-source code and patents lie further apart than the randomization benchmark places them. The excess, averaged over a firm's years, is positive in 97.1% of the 595 firms. It is positive in every year for 93.3% of firms.

The composition of the excess depends on firm-year weights and the level of patent classes. With each firm-year counted once, open-source code outside the patent support accounts for 85.7% of the mean excess under event weights and 63% under equal weights (Table A1). A firm's *open-source program* is the set of open-source projects it runs in a firm-year. When I weight each firm-year by its open-source events, so that the largest programs count most, open-source code outside the patent support accounts for 6.14% under event weights. At coarser levels of the CPC, with each firm-year counted once, it accounts for 22.3% at the subclass level (such as G06F) and 12.9% at the class level

(such as G06). Most of the typical firm-year's excess therefore lies outside the patent support. In the largest programs and at coarser levels, most lies inside the patent support, where open-source code and patents differ in their shares across patent classes. The split also depends on the labels. With the labels of a different model that chooses main groups within the same subclasses as the production model, open-source code outside the patent support accounts for 97.3% of the mean excess under event weights, with each firm-year counted once, against 85.7% with the production labels. The excess itself keeps 96% of its size (Appendix C).

4.4 What firms open instead

What firms open instead is mostly software engineering and computing, often outside the technologies they patent. Section 3.4 shows where the firms' open-source code sits. The randomization benchmark reallocates technologies among a firm's patents and open-source projects. Under the randomization benchmark, a firm's open-source code therefore follows, on average, the mix of all the firm's items, patents included. Averaged over the comparison sample's 2,472 firm-years under event weights, software engineering (G06F 8/00) accounts for 22.09% of a firm's open-source code, against 10.64% under the randomization benchmark, and program control (G06F 9/00) accounts for 11.99%, against 6.64%. The main groups other than the 15 with the largest observed shares account for 27.81% of the firm's open-source code, against 52.61% under the randomization benchmark. Relative to the randomization benchmark, firms' open-source code thus has larger shares in software engineering and program control and smaller shares in other technologies.

Technology norms reproduce about half of the excess. Under event weights, technology norms reproduce 49% of the excess (95 percent interval 46.7% to 51.3%). A stricter redraw at the technology's rate among the other firms in the same industry adds a further 0.7%, with a 95 percent interval from −0.5% to 1.8% that includes zero (Appendix B.3). The rest, 50.3% (47.9% to 52.7%), is the *remainder*. In the remainder, the firm's open-source code and patents are further apart than technology norms and the stricter redraw imply (Table 5). Under equal weights, technology norms reproduce 51.3% and the remainder is 48.3%. Technology norms reproduce all of the mismatch inside the patent support, and more (136.9%), but only 34.3% of the open-source code outside it, so the remainder lies mostly outside the patent support (Appendix Table B7). The excess is consistent with earlier work arguing that patents record only part of what a firm knows (Jaffe, 1989, p. 89, Podolny et al., 1996, p. 672). Within a firm, research disclosures also differ from its own patents. They cover older or discontinued lines of research (Ganglmair and Kann, 2025a, slides 18 and 20). In economic terms, half of the excess is what any firm holding the same technologies would show, and the other half sets the firm apart from its technology peers.

The excess can also arise without changes in a firm's portfolios over time. In the fixed-portfolio world, where each firm keeps its patents and open-source code over the whole period and only the year of each item is redrawn, the excess is 0.2373 under event weights, against an observed 0.2336 (Appendix B.3).

Table 5: The excess reproduced by technology norms

	Equal weights	Event weights
Panel A. The excess over the randomization benchmark		
Observed	0.3304	0.2336
Panel B. The mean distance		
Observed	0.7645	0.8055
The randomization benchmark	0.434	0.5719
The randomization benchmark at technology norms	0.6035	0.6864
Panel C. The share of the excess, with its 95 percent interval		
Technology norms	51.3% [49.4%, 53.1%]	49% [46.7%, 51.3%]
The remainder	48.3% [46.4%, 50.3%]	50.3% [47.9%, 52.7%]

Notes: Means over the comparison sample's 2,472 firm-years of 595 firms. Panel A reports the excess, which is the observed distance minus the benchmark distance. The benchmark distance is the mean distance over the benchmark's 200 draws. Panels B and C redraw the randomization benchmark. As in the randomization benchmark, each item's set of main groups moves together, and each start year keeps its number of open-source projects. At technology norms, the draw selects which items are open-source projects, at odds set by the rate at which the other firms put each item's main groups into open-source code. In Panel C, technology norms are the randomization benchmark at technology norms minus the randomization benchmark. The remainder is the observed distance minus a stricter redraw at the rate among the other firms in the same industry, which adds 0.4% under equal weights and 0.7% under event weights beyond technology norms (Appendix B.3). Each component is a share of the observed excess, with a 95 percent interval clustered by firm.

The separation might imply choice, though it need not. The remainder is net of the firm's own counts and technology norms, and it is what a firm would produce by choosing what to put into open-source code technology by technology. Consider a model in which, in each technology, a firm chooses how to allocate each invention to secrecy, a patent, open-source code or both. In that model, a choice that ignores technology gives no excess over the randomization benchmark, whatever the firm's size or overall openness, while a choice that depends on technology gives one. A technology's norm is the part of that choice the firm shares with the technology's other holders, and these norms give the part of the excess that the decomposition's technology effects reproduce. The firm's own *deviation* in a technology is what remains beyond the norm and the firm's overall openness. In that model, the firm chooses its deviations, and they give the remainder. The results are consistent with this interpretation. The excess is present in nearly every firm-year, and technology norms reproduce about half of it. Yet unchosen differences between firms would leave the same remainder. One is the kind of open-source code a firm writes. Another is the breadth of its items. On average, a patent carries 3.01 main groups and an open-source project carries 1.78 (Appendix A.7). In Appendix E.8, tests of stated reasons for publishing open-source code do not single out any one reason.

5 The excess and the size of open-source programs

Section 4 answers the paper's question for the average firm. That firm's open-source code reaches into technologies it does not patent, and in nearly every firm the open-source code lies further from the patents than the randomization benchmark built from the firm's own items places it (section 4.3). This section asks whether that separation reflects each firm's technology mix or the size of its open-source program. If what a firm opens did not depend on technology, its open-source code and its patents would be draws from one pool. The distance between them would then fall with the count of open-source projects at the benchmark's pace. Panel C of Table 7 reports the excess, which nets out the benchmark distance at each firm's own counts, and it is the panel of interest. Section 5.1 sets out what the randomization benchmark predicts, section 5.2 reports the regressions, section 5.3 describes how firms differ, section 5.4 reads two arguments in cited sources, section 5.5 compares each firm with its industry, and section 5.6 summarizes.

5.1 What the randomization benchmark predicts

I measure the size of a firm's open-source program by its count of open-source projects. This count does not measure the firm's intent. Over five years, the median firm-year in the sample has 35.0 open-source projects. The 10th percentile is 7.0 and the 90th percentile is 312.4. Firms with a handful of projects participate on a small scale. Those with hundreds run an open-source program. Firms in the comparison sample are large, and they both patent and publish open-source code. I take the size of each firm's program as given and ask only which technologies it covers.

Assumption 1. A firm's numbers of open-source projects and patents are set apart from which technologies those items carry.

The randomization benchmark rests on Assumption 1. It holds fixed each firm's counts and the weights of its items, and it reassigns only their technologies (section 4.2). The data provide one check on the assumption. With a firm's counts held fixed, its excess shows no clear association with its size or its R&D intensity (section 5.3).

If what a firm opens did not depend on technology, its open-source code and its patents would be draws from one pool, the technologies of the firm's own items. As the counts grow, the shares of both portfolios approach the shares of that pool, by the law of large numbers, so the distance between them falls. A distance between two small portfolios is positive even when their items are allocated at random (section 4.2). Carrington and Troske (1997, p. 402) call this the small-unit bias of dissimilarity indices, and it shrinks as the counts grow (Appendix A.8).

Prediction 1. Under the randomization benchmark, the distance between a firm's open-source code and its patents falls as the firm's count of open-source projects grows.

The benchmark distance falls gradually with the number of items. In the third of firm-years with the most open-source projects, the benchmark distance computed from equal numbers of items

from each channel falls from 0.810 with three of each to 0.431 with 320 of each, while the observed distance falls only from 0.881 to 0.712.[7] Prior work documents the decline in distance as the count grows. Benner and Waldfogel (2008, NBER Working Paper 13322, p. 15) draw equal numbers of patent class listings per firm to show how proximity depends on the count, and D'Haultfœuille and Rathelot (2017, p. 62) show that even an index corrected against random allocation depends on the size of its units. This section adds a reference for the count slope, the benchmark distance at each firm's own counts.

To compare the count slope for the observed distance with that for the benchmark distance, I estimate

$$\begin{aligned} Y_{it} = \beta_R \log(1 + n^{\text{oss}}_{i,t-1}) + \beta_E \log\Big(1 + \frac{n^{\text{ev}}_{i,t-1}}{n^{\text{oss}}_{i,t-1}}\Big) + \beta_P \log(1 + n^{\text{pat}}_{i,t-1}) \\ + \delta\, \psi_{it} + \phi_{it} + \varepsilon_{it}, \qquad Y_{it} \in \{D_{it},\ \mu_{it},\ \tilde{D}_{it}\}. \end{aligned} \tag{1}$$

Under event weights, the outcome Y_{it} is, in turn, the observed distance D_{it} between firm i's open-source code and its patents over the five years ending in year t, the benchmark distance μ_{it}, which is the mean over the benchmark's 200 draws, and the excess $\tilde{D}_{it} = D_{it} - \mu_{it}$. Because distance varies with portfolio size, I estimate the same specification for all three outcomes. The regressors $n^{\text{oss}}_{i,t-1}$, $n^{\text{ev}}_{i,t-1}$ and $n^{\text{pat}}_{i,t-1}$ count the firm's open-source projects, open-source events and patents over the five-year window that ends one year earlier. Each enters as the log of one plus the count, which keeps firm-years with no items of a given type in the sample. Events enter per open-source project, so that β_E measures how active the firm's projects are. The coefficient of interest is β_R. The term ψ_{it} is the share of the firm's patents outside the classes that carry open-source code, with coefficient δ. The term ϕ_{it} denotes fixed effects, and ε_{it} is an error term. With industry-by-year fixed effects, β_R compares firms in the same industry and year. With firm fixed effects, it compares a firm with itself over time. Because the count and outcome windows share four years, β_R measures an association between overlapping histories, not a response to earlier activity. I cluster standard errors by firm (Cameron and Miller, 2015). Let β^D_R, β^μ_R and $\beta^{\tilde{D}}_R$ denote β_R when the outcome is the observed distance, the benchmark distance and the excess. The three regressions use the same sample and regressors, so $\beta^D_R = \beta^\mu_R + \beta^{\tilde{D}}_R$. Prediction 1 says that β^μ_R is negative.

The null hypothesis H_0 is that the observed distance falls with the count at the benchmark's pace, $\beta^D_R = \beta^\mu_R$, or equivalently $\beta^{\tilde{D}}_R = 0$. The alternative is two-sided. I test H_0 in two ways. The first uses the coefficient on log open-source projects when the excess is the outcome, with its standard error clustered by firm. The second uses randomization inference (Young, 2019, pp. 572–573). I draw the randomization benchmark 1,000 more times, recompute every firm-year's distance in each draw, and re-estimate the regression using these distances with the same regressors, fixed effects and firm-years. In the regression across firms, I also recompute ψ_{it} in each draw, because the

[7] For each firm-year of the comparison sample, under event weights, I select the same number of the firm's open-source projects and of its patents at random, from three of each up to 320 of each, 200 times. For each selection, I compute the distance and the benchmark distance, with the randomization benchmark reassigning the selected items' technologies among them without start-year strata. The thirds split the firm-years by their count of open-source projects within each year. Of the 819 firm-years in the top third, 157 have 320 items in each channel.

randomization benchmark changes it. The randomization p-value compares the observed coefficient with the coefficients from the draws.

Rejecting H_0 shows that a firm's open-source code is not a random draw from its own technologies at each count. It does not show that firms with larger programs keep their open-source code further from their patents. The data would also reject H_0 if each firm's open-source code kept one technology mix of its own at every size. Open-source code whose mix differs from the patents' mix responds less to the small-unit bias than open-source code reallocated at random, so the excess rises with the count whenever the two keep different mixes (Appendix B.12). Larger programs are also measured with more items. The median firm-year holds 11 open-source projects in the third with the fewest, 42 in the middle third and 211 in the third with the most. Simulated firms that keep one technology mix of their own at every size show how large that rise is (section 5.2). Table 6 sets out what each case predicts.

Table 6: What each case predicts for the excess

Case	Slope of the excess on the count, at each firm's own counts
1. The open-source code is a random draw from the firm's own technologies (H_0)	Zero
2. The open-source code keeps a technology mix of the firm's own at every size	Positive, as in simulated firms that keep their own mix
3. The open-source code lies further from the patents, relative to the randomization benchmark, as the program grows	Positive, and larger than in simulated firms that keep their own mix

Notes: The slope is $\beta_R^{\tilde{D}}$ in equation (1), the coefficient on log open-source projects with the excess as the outcome, at each firm-year's own counts. The simulated firms keep each firm's own counts, technology mix and patents at every size (Appendix B.12). Section 5.2 reports the slope and the simulation.

The same reasoning applies when I compare the firm with its industry (section 5.5). The industry-pool benchmark draws each of the firm's portfolios, at its own size, from its channel's pool in the industry. Under it, each of the firm's portfolios comes nearer its industry's portfolios as the firm's count in that channel grows, again through the small-unit bias. The test is the same, with the excess between a firm's portfolio and its industry's as the outcome.

5.2 More open-source code and the excess

5.2.1 The regression estimates

Table 7 reports the estimates, with the three outcomes in Panels A, B and C. All seven columns use every firm-year of the comparison sample that the fixed effects leave identified. They differ in their fixed effects, from none in column 1 to industry-by-year effects in column 4 and firm effects in columns 5 to 7. The columns without firm effects, 1 to 4, also include ψ_{it}. I focus on two columns.

Column 4 compares firms in the same industry-year. Column 6 compares a firm with itself over time, with firm and year fixed effects. Because the three regressions in a column use the same sample and regressors, each coefficient in Panel A equals the sum of the corresponding coefficients in Panels B and C. Appendix Table B1 repeats columns 4 and 6.

Firms that publish more open-source code have it nearer their patents, but the randomization benchmark at their own counts brings it nearer faster. Panel A of Table 7 reports the estimates for the observed distance. The coefficient on log open-source projects is negative in every column. Across firms in the same industry-year (column 4), a doubling of the open-source projects is associated with a change of -0.026 in the observed distance. Within a firm over time (column 6), a doubling is associated with a change of -0.010.

Table 7: The distance, the benchmark distance and the excess, and the firm's counts

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Observed distance							
Log open-source projects	−0.041***	−0.042***	−0.037***	−0.038***	−0.017***	−0.015***	−0.016***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Log events per project	0.00046	0.00033	0.004	0.004	0.006***	0.006***	0.006***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Log patents	0.005**	0.005**	0.00033	0.00039	−0.002	−0.001	−0.001
	(0.002)	(0.002)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Patent share outside	0.280***	0.279***	0.250***	0.251***			
open-source classes	(0.017)	(0.017)	(0.020)	(0.021)			
Panel B. Benchmark distance (mean over the benchmark's 200 draws)							
Log open-source projects	−0.076***	−0.077***	−0.071***	−0.071***	−0.049***	−0.066***	−0.060***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)
Log events per project	0.015***	0.015***	0.019***	0.019***	0.014***	0.012***	0.016***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log patents	0.004*	0.004*	−0.005	−0.005	−0.002	−0.011**	−0.011
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.006)	(0.007)
Patent share outside	0.198***	0.196***	0.174***	0.177***			
open-source classes	(0.022)	(0.023)	(0.026)	(0.027)			
Panel C (the panel of interest). Excess over the randomization benchmark (observed minus benchmark distance)							
Log open-source projects	0.035***	0.035***	0.034***	0.034***	0.032***	0.051***	0.044***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.007)	(0.007)
Same, without	0.030***	0.030***	0.032***	0.032***			
the patent share	(0.004)	(0.004)	(0.004)	(0.004)			
Log events per project	−0.014***	−0.014***	−0.015***	−0.015***	−0.008	−0.006	−0.010*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Log patents	0.002	0.002	0.005	0.005	0.00018	0.009	0.010
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.007)
Patent share outside	0.081***	0.082***	0.076***	0.074***			
open-source classes	(0.021)	(0.021)	(0.025)	(0.026)			
Firm fixed effects	No	No	No	No	Yes	Yes	Yes
Year fixed effects	No	Yes	No	No	No	Yes	No
Industry fixed effects	No	No	Yes	No	No	No	No
Industry-by-year fixed effects	No	No	No	Yes	No	No	Yes
Observations	2,447	2,447	2,447	2,447	2,447	2,447	2,447
Firms	589	589	589	589	589	589	589

Notes: The table reports estimates of equation (1), with the distances under event weights. The outcome is the observed distance in Panel A, the benchmark distance (the mean distance over the benchmark's draws) in Panel B, and the excess (the observed distance minus the benchmark distance) in Panel C. Panel C is the panel of interest, because the excess nets out the benchmark distance at each firm-year's own counts. The three panels of a column use the same sample and regressors, so each coefficient in Panel C equals the one in Panel A minus the one in Panel B. Each count is over the five-year window that ends the year before the outcome's window ends, and it enters as the log of one plus the count. A coefficient is thus the change in the outcome per unit of that log, and the text multiplies the coefficients in columns 4 and 6 by $\ln 2$, approximately the change when the count doubles. All columns use every firm-year of the comparison sample that the fixed effects of column 7 leave identified. Columns 1 to 4 also include the patent share outside open-source classes, the share of the firm's patents outside the main groups that carry some firm's open-source code, in the outcome's window. The row "Same, without the patent share" gives the coefficient on log open-source projects when that share is left out. The industry is the firm's four-digit NAICS industry, or its three-digit industry when fewer than five firms have linked open-source code. Panels A and B omit the row without the patent share. Appendix B.9 reports it and repeats the regressions for the 883 firm-years that also have the leader indicator, product-market closeness, eligibility exposure and first-time outside participants. Standard errors clustered by firm are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from standard errors clustered by firm.

Panel B reports the same regressions for the benchmark distance. Consistent with Prediction 1, the coefficient on log open-source projects is negative in every column, and it is larger in absolute value than in Panel A. A doubling of the open-source projects is associated with a change in the benchmark distance of −0.049 across firms in an industry-year and of −0.046 within a firm. The coefficients in Panel A therefore combine the small-unit bias with any change in how the firm's

open-source code and its patents differ. Panel C separates the two.

Panel C reports the estimates for the excess, the first test of H_0. At each firm's own counts, the excess is higher in larger programs, and its coefficient on log open-source projects is positive in every column. A doubling of the open-source projects is associated with a change of 0.023 in the excess across firms in an industry-year (column 4) and of 0.036 within a firm (column 6). The observed distance therefore falls far more slowly than the benchmark distance as the count rises, and H_0 is rejected in every column.

5.2.2 Randomization inference for the count slope

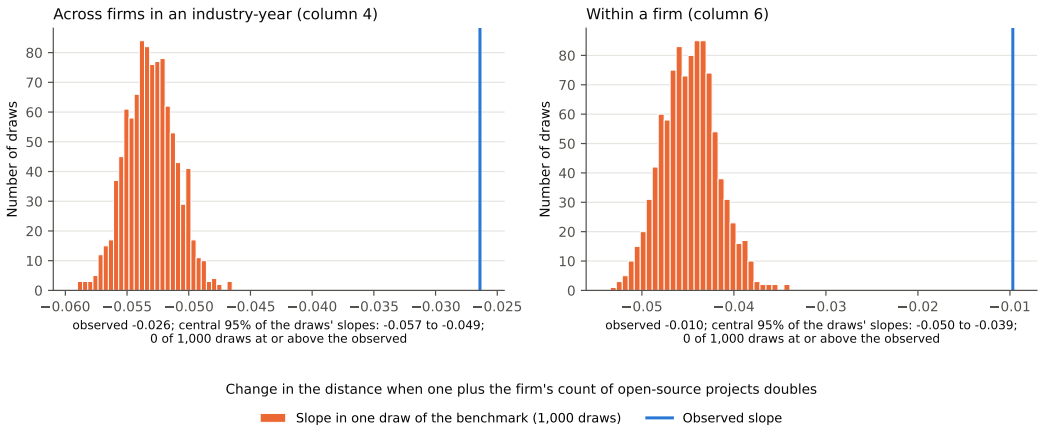


Figure 3: Randomization inference for the count slope

Notes: Each panel compares the observed count slope with the slopes from the randomization benchmark. The count slope is the coefficient on the log of one plus the firm's open-source projects over the five years ending in the preceding year, times $\ln 2$, which is approximately the change in the distance D_{it} between the firm's open-source code and its patents, under event weights, when that count doubles. Each draw of the randomization benchmark reassigns the technologies of each firm's own patents and open-source projects at random among items with the same start year (the year an item is first seen), recomputes every firm-year's distance and re-estimates the regression with the same regressors and fixed effects. Each bar counts the draws whose slope falls in its bin, out of 1,000 draws, and the vertical line is the observed slope. The bars show what the randomization benchmark produces, not uncertainty around the observed slope. Left: across firms in an industry-year (Table 7, column 4), with industry-by-year effects and the share of the firm's patents outside the classes that carry open-source code, computed again in each draw. The observed slope is -0.026, and the central 95 percent of the draws' slopes lie between -0.054 and -0.046. Right: within a firm (column 6), with firm and year effects. The observed slope is -0.010, and the central 95 percent of the draws' slopes lie between -0.053 and -0.040. The number of draws at or above the observed slope is 0 on the left and 0 on the right, so the two-sided randomization p-value is 0.0020 on the left and 0.0020 on the right, the smallest that 1,000 draws allow. The sample is the 2,447 firm-years of 589 firms retained in the regressions after singleton observations are dropped for the fixed effects.

Randomization inference, the second test of H_0 (section 5.1), also rejects it (Figure 3). Across firms in an industry-year, the observed coefficient per doubling is -0.026, and the central 95 percent of the coefficients from the 1,000 draws lie between -0.054 and -0.046. Within a firm, the observed coefficient per doubling is -0.010, and the central 95 percent of the coefficients from the draws lie

between −0.053 and −0.040. No draw yields a coefficient as close to zero as the observed one in either column. The two-sided randomization p-value is therefore 0.0020 across firms and 0.0020 within a firm (one-sided, 0.0010 and 0.0010), the smallest attainable with 1,000 draws. The sample is the 2,447 firm-years of 589 firms that remain after singleton observations are dropped for the fixed effects.

5.2.3 Simulated firms that keep a technology mix of their own

Rejecting H_0 does not make firms of different sizes comparable (section 5.1). In simulations that use each firm's own counts, mix and patents and keep its technology mix fixed at every size, a doubling of the open-source projects is associated with a change of 0.0221 in the excess across firms and of 0.0324 within a firm, compared with 0.023 and 0.036 in Panel C (Appendix B.12). The data are thus consistent with each firm keeping one technology mix of its own at every size, the second case of Table 6. Two simulated worlds bracket this reading in every column of Table 7. Their slopes below are per unit of the log count, as the table reports them, rather than per doubling. In those units, the data's slope of the excess is 0.0337 in column 4 and 0.0515 in column 6. Each world is simulated many times, and its band holds the central 95 percent of the slopes over these replicates. In a placebo world, each open-source project takes the technologies of one of the firm's own patents, so the two channels share one mix. The slope of the excess is then near zero in every column. In column 4, its mean over the replicates is −0.003, with a band from −0.0075 to 0.0006. No replicate reaches the data's slope in any column, so each Monte Carlo p is 0.01, the smallest the replicates allow. Holm's adjustment across the seven columns, with and without the patent share outside open-source classes, raises this floor to 0.139. That value reflects the number of replicates rather than how far the data lie from the placebo world. In the world in which each firm keeps one mix at every size, the data's slope of the excess lies within the world's band in every column. The world's slope is 0.0319 in column 4 and 0.0468 in column 6 (Monte Carlo $p = 0.299$ and 0.169). Under a fixed separation between the two mixes, the measure alone thus produces a slope of the observed size. The data do not show that the mix changes as the firm publishes more. The estimates in Panel C therefore cannot establish whether firms with larger open-source programs keep their open-source code further from their patents than firms with smaller ones.

In the same placebo world, three excesses between the firm and its industry have a positive slope on the firm's count in the channel each distance starts from. They are the excesses from the firm's open-source code to its industry's open-source code and to its industry's patents, on the repository count, and from the firm's patents to its industry's open-source code, on the patent count. The world's 95 percent band lies above zero on all twelve primary cells, columns 4 and 7 of the regressions in section 5.5, each with and without the patent share outside open-source classes. For the excess from the firm's open-source code to its industry's open-source code, within a firm, the world's slope is 0.0373, with a band from 0.0324 to 0.0432, against 0.0176 in the data, per unit of the log count. On every primary cell the data's slope lies below every replicate, and Holm's adjustment across the cells raises the Monte Carlo p to 0.119, the floor the replicates allow. The

sign of these excess slopes is thus a property of the measure at the firm's counts. In this world the firm's code takes patent-like technology vectors, and firms without patents keep their own code.

5.2.4 Robustness

At the firm's own counts, the positive coefficient on log open-source projects in the excess regression does not come from patents outside the classes that carry open-source code. Such patents lie in classes with no open-source code, and their share is a lower bound on the distance (Appendix A). The coefficient remains positive after I control for ψ_{it}. Without fixed effects (column 1 of Table 7), adding ψ_{it} changes the coefficient on log open-source projects in Panel C from 0.030 to 0.035. With industry-by-year fixed effects (column 4), it changes the coefficient from 0.032 to 0.034. Multiplied by $\ln 2$, the latter is the coefficient of 0.023 per doubling reported above.

Restricting both portfolios to the classes that carry open-source code preserves the result (Appendix Table B8). For those classes alone, with sector and year fixed effects, the coefficient on log open-source projects in the excess regression is 0.0345 (standard error 0.0044).[8] Within these classes, the observed distance falls as the program grows, but more slowly than the benchmark distance. At each firm's own counts, the positive coefficient on log open-source projects in the excess regression persists across windows and fixed effects, except for the three-year blocks. It is positive in all seven columns of Table 7, and it ranges from 0.032 with firm fixed effects alone to 0.051 with firm and year fixed effects. With firm and year fixed effects and event weights, the coefficient is 0.0516 for five-year windows, 0.0527 for three-year windows and 0.03 for single years (Appendix Table B5). When I compare each three-year block with the preceding block, the coefficient is 0.0164 (standard error 0.0112) and is not statistically different from zero. Over five-year blocks, it is 0.0833 in long differences and 0.0188 in the cross-section. Appendix B.9 reports every coefficient and repeats the regressions for the 883 firm-years that also have the leader indicator, product-market closeness, eligibility exposure and first-time outside participants.

5.3 How firms differ

Firms may differ in what they reveal, and Henkel (2006, p. 953) finds revealing strongly heterogeneous among firms. No cited source states how the excess should differ by firm type, and the comparisons below are descriptive. Across firms, the mean excess ranges from 0.066 at the tenth percentile to 0.390 at the ninetieth percentile.

At each firm's own counts, larger firms and more research-intensive firms show no clear difference in the excess, consistent with Assumption 1. I specified four comparisons before testing. They

[8] Outside the classes that carry open-source code, a firm's patents lie in different main groups from any open-source code, so part of its distance could come from them alone. I therefore keep only the firm's patents in the classes that carry open-source code, rescale their shares to sum to one, and measure their distance to the firm's open-source code, D_{OSS}, whose subscript stands for the classes that carry open-source code. All of the open-source code lies in those classes. The regression controls for ψ_{it} and includes sector and year fixed effects, where a *sector* is a two-digit NAICS code. In it, the coefficient on log open-source projects is -0.0399 for D_{OSS} and -0.0745 for its benchmark distance. The coefficient in the excess regression is the first minus the second.

concern the count of open-source projects, R&D intensity, size (log total assets) and industry. Because the benchmark distance can vary with each of these characteristics, I compare each observed coefficient with the corresponding coefficient re-estimated for each draw of the randomization benchmark, and I adjust the four p-values with Holm's method. This randomization inference follows Young (2019, pp. 572–573), Duranton and Overman (2005, LSE Research Online version, p. 10), Gentzkow et al. (2019b, p. 1318) and Carrington and Troske (1997, pp. 405–406). Under event weights, on the 1,743 firm-years with total assets in the size comparison, the excess's coefficient on log total assets is 0.0018 (standard error 0.0033), and that on R&D intensity is 0.0688 (standard error 0.0601). Neither survives Holm's adjustment. The observed coefficient on log open-source projects exceeds its mean over the benchmark's draws by 0.0516 (standard error 0.0072). Appendix B.11 gives the equation with its terms, Holm's adjustment and all four results (Table B41).

Larger open-source programs show a larger excess at their own counts, but not at equal numbers of items. At each firm's own counts, the excess rises from 0.1855 in the quarter of firm-years with the fewest open-source projects to 0.2762 in the quarter with the most (Appendix Table B24). Simulated firms that keep one technology mix at every size show a rise of this kind (section 5.2.3). To compare programs of different sizes, I pick five open-source projects and five patents at random from each firm-year's window, 200 times. On each pick, I compute the distance and the benchmark distance, with the benchmark reassigning the technologies among the ten picked items. At these equal numbers of items, a doubling of the open-source projects goes with a change of -0.007 (standard error 0.002) in the excess across firms in an industry-year and of 0.00023 (standard error 0.002) within a firm. Once the comparison holds the number of items fixed, the excess does not rise with the size of the program within a firm, and across firms it falls slightly.

Beyond the four declared comparisons, I compare firms by seniority, with indicators that the panel truncates. A firm counts as senior in the panel when it is in the covariate file from 2010, the file's first year, and as senior in open source when its first open-source year is 2015, the first year of the open-source events. Both years are floors, so each indicator puts the oldest firms together with firms that entered at the floor. The indicators do not vary within a firm, so these comparisons include no fixed effects. As in the declared comparisons, I compare each observed coefficient with the coefficient re-estimated in each draw of the randomization benchmark, and I adjust the p-values with Holm's method over one family that holds size and the two indicators. For firms senior in the panel, the open-source code lies further from the patents, relative to later entrants, than the benchmark's draws produce. Their coefficient is 0.0578, against 0.0348 in the draws (Holm-adjusted $p = 0.015$). For firms senior in open source, it lies nearer, with a coefficient of -0.0238 against -0.001 in the draws (Holm-adjusted $p = 0.02$). With standard errors clustered by firm, neither difference in the excess is distinguishable from zero after Holm's adjustment ($p = 0.178$ and 0.178). In the same family, size lies within the draws on the 2,318 firm-years with total assets, with a difference of -0.002 (standard error 0.0023).

Of the three declared comparisons beside the count slope, industry is the one associated with the excess. The excess differs across industries, from 0.166 in aerospace products to 0.385 in

pharmaceuticals, and the mean excess is positive in every industry with at least five firms. In the industries with the largest excess, it comes mostly from the technologies the industries patent. Technology norms reproduce 80% of the pharmaceutical firms' mean excess and 60% of the instrument firms', against 49% of the excess as a whole. Beyond technology norms, the pharmaceutical firms' mean excess falls to 0.078, among the smallest industry means. Beyond technology norms, industry differences remain but are smaller. The standard deviation of the industries' mean excess falls from 0.038 to 0.025. Industry differences remain when the count of open-source projects is held fixed.

Most of the open-source code outside the patent support lies in technologies related to the firm's patented ones, though the share is slightly lower than under the randomization benchmark. The distance compares the full portfolios. I also examine the firm's open-source code outside the patent support, main group by main group, and ask whether each main group is related to a main group in the firm's patents. Two main groups count as related when at least five US patents granted in 2015–2024 carry both, and when that count lies at least three standard deviations above its mean under a random assignment of the main groups to patents within each grant year. In the average firm-year, under event weights, 90% of the open-source code outside the patent support (section 4.3) lies in main groups related to a patented one, against 95.9% under the randomization benchmark and 86.2% under technology norms. Averaged over firm-years, the excess outside the patent support is 20% of the open-source code, and it lies mostly in related main groups (17.5%) and partly in unrelated ones (2.5%). When I instead weight each firm-year by its open-source events, so that the largest programs count most, the related share is 69.3%, against 94% under the randomization benchmark. In the larger programs, the open-source code outside the patent support leans further toward unrelated technologies.

A small share of the open-source code enters technologies new to the firm. Of the weight of the open-source projects firms start in 2020–2024, 3.3% falls in main groups where the firm held neither a patent filing nor an active project in the five previous years. Patent filings are complete through 2020, and of the patents filed in that year in the classes that carry open-source code, 1.6% falls in such main groups. Each share lies above the band of a timing benchmark that permutes the dates of the firm's own items, from 2.7% to 2.8% for projects and from 1.2% to 1.3% for patents. Each channel thus enters new main groups slightly more often than permuted timing places it there, yet both stay almost entirely inside main groups the firm already held. Each channel is read against its own band rather than against the other channel, and the projects' events start in 2015, so earlier code is unseen.

Firms are also more likely to file a patent within two years in a main group they enter through open-source code than in main groups they do not enter that year (coefficient 0.043, standard error 0.007). When the firm's code entry years are exchanged across its own technologies, or a technology's entry years across the firms that entered it, the filing lies as close to the code as in the data (Holm-adjusted $p = 0.952$ and two-sided $p = 0.222$). The association thus lies in which technologies a firm codes and patents in rather than in a patent arriving soon after the code, as far as each reference's form of common timing shows. Between firms in the same industry and year, a

firm that enters a technology through open-source code files a patent there within two years more often than other firms in its industry and year file in technologies that are new to them and that they did not enter (coefficient 0.046, standard error 0.007; within four years, 0.067, standard error 0.017). With the entry years exchanged in the same two ways, the filing again lies as close to the code as in the data (two-sided $p = 0.878$ and 0.164). This is an association between firms.

5.4 Tests against stated arguments

Two arguments in cited sources, a recommendation and a finding, can be read in these data against the randomization benchmark. Table 8 reports them. When the co-creators of value are distributed, numerous and anonymous, Henkel et al. (2013, p. 72) recommend open modules beside protected ones (their Table 1, p. 70). A repository that draws more outside participants stands in here for that condition. Panel A regresses the share of a repository's code outside the firm's patent support over a five-year window on the log of one plus its outside participant-years, with its events and age as controls. Outside participant-years count, year by year, the people active on the repository who were not among its first contributors and had no earlier event on the firm's accounts. Teece et al. (1994, p. 9) find that firms keep a constant level of coherence between neighboring activities as they grow more diverse. Panel B reads that finding with the related share, the share of the firm's open-source code outside the patent support that lies in main groups related to a patented one (section 5.3), and asks whether it changes with the breadth of the firm's patents. In each panel, I re-estimate the regression on each draw of the randomization benchmark and compare the observed coefficient with the draws' mean and band.

Across a firm's repositories in the same year, those with more outside participant-years keep a slightly smaller share of their code outside the firm's patented technologies than the randomization benchmark gives (Table 8, column 1). The coefficient is -0.0039, against 0.0001 over the benchmark's draws, a difference of -0.004. Within the firm-year, the difference changes little when the length of the text the language model read enters as a control (column 3), so thin text does not account for it. The test reads a placement only when both of its p-values lie below 0.05. The Monte Carlo p places the observed coefficient among the coefficients over the benchmark's draws, and the regression p also counts the coefficient's sampling error, clustered by firm. Between firms in the same industry and year, the difference is -0.0015 (column 2). The coefficient lies outside the central 95 percent of the benchmark's draws (Monte Carlo $p = 0.02$). Once its sampling error is counted, the regression p is 0.399, and the column is thus not distinguishable from the benchmark. With the length of the text held fixed, the difference between firms is -0.0055 (column 4, Monte Carlo $p = 0.01$, regression $p = 0.029$). When each repository-window is weighted by its events, the difference is -0.0028 and is not distinguishable from zero (regression $p = 0.111$). The placement is thus not consistent with the separation that Henkel et al. (2013) recommend. Outside participant-years also count new hires, and the test describes where outside participants work, not why. In Panel B, the coefficient on the breadth of the patents lies inside the band of the benchmark's draws, so the related share's association with breadth is not distinguishable from what the benchmark gives. This is consistent

Table 8: Tests against stated arguments

	(1)	(2)	(3)	(4)
Panel A. Outside participants and the share of a repository's code outside the patent support				
Log outside participant-years	−0.0039	0.0000	−0.0048	−0.0056
	(0.0011)	(0.0027)	(0.0013)	(0.0039)
Mean over the benchmark's draws	0.0001	0.0015	−0.0001	−0.0001
Difference from the draws' mean	−0.004***	−0.0015	−0.0046***	−0.0055**
	(0.0011)	(0.0018)	(0.0012)	(0.0025)
Monte Carlo p	0.01	0.02	0.01	0.01
Log length of the text read			Yes	Yes
Repository-windows	473,051	473,051	266,609	266,638
Panel B. The breadth of the firm's patents and the related share of its code outside the patent support				
Log breadth of the patents		0.0212		
		[−0.0217, 0.0642]		
Draws' mean [band]		0.0115 [−0.0091, 0.0339]		
Software share of the patents		0.107		
		[0.0193, 0.1947]		
Log patents		−0.0349		
		[−0.06, −0.0099]		
Firm-years (firms)		2,354 (581)		
Firm-year effects	Yes		Yes	
Industry and year effects		Yes		Yes

Notes: Panel A: a repository-window is a repository's five-year window under a firm. The outcome is the share of its code outside the firm's patent support in that window, and the regressor is the log of one plus its outside participant-years, with the log of its events and its age as controls. Columns 1 and 3 compare a firm's repositories in the same year, with firm-year effects. Columns 2 and 4 compare firms in the same industry and year, with industry and year effects only, the industry as in Table 7. Columns 3 and 4 add the log length of the text the language model read for the repository, on the repository-windows for which that length is recorded. The mean over the benchmark's draws re-estimates the coefficient with each draw of the randomization benchmark. The difference is the observed coefficient minus that mean. Parentheses under a coefficient or a difference give its standard error clustered by firm. The Monte Carlo p is two-sided and places the observed coefficient among the coefficients over the draws. The stars and the text read the difference against its standard error clustered by firm, the regression p. Panel B: the outcome is the share of the firm-year's open-source code outside the patent support that lies in main groups related to a patented one. It is regressed on the log number of main groups in the firm's patents (breadth), the share of its patents in software classes, log patents and log open-source events, with two-digit NAICS and year effects. Brackets under a coefficient give its 95 percent interval clustered by firm, and the band gives the central 95 percent of the coefficients over the benchmark's draws. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, for the difference from standard errors clustered by firm.

with the finding of Teece et al. (1994, p. 10) that coherence changes little as firms grow more diverse. The reading is weak, because the interval clustered by firm is wide, and breadth moves closely with the patent count (correlation 0.92). The coefficients on the software share and on log patents describe the sample and test no cited statement.

5.5 Extension: the firm beside its industry

The firm's two portfolios and its industry's two define six distances. The *industry's portfolios* combine the patents and the open-source code of the other firms in its industry, excluding the firm itself. The industry is the firm's four-digit NAICS industry in the previous year, or its three-digit industry when the four-digit industry contains fewer than five firms with open-source code. Let the firm's patents and open-source code be P_{it} and O_{it}, and the industry's be $P_{G,t}$ and $O_{G,t}$. The firm's own distance is $D_{it} = D(P_{it}, O_{it})$, and the industry's own distance is $D(P_{G,t}, O_{G,t})$. The remaining four distances compare the firm with its industry. They are $D(O_{it}, O_{G,t})$, $D(O_{it}, P_{G,t})$, $D(P_{it}, O_{G,t})$ and $D(P_{it}, P_{G,t})$. Because the distance is a metric, the six distances bound one another. By the triangle inequality, the firm's open-source code and its patents can differ in their distance to either of the industry's portfolios by at most the firm's own distance. Appendix Table G10 reports the six distances for the 2,450 firm-years of 594 firms with an industry.

Two randomization benchmarks provide references for the four distances between the firm and its industry. The *industry-pool benchmark* reassigns the technologies of the industry-year's items within each channel and start year, including the firm's own items. It assigns as many items in each channel as the firm holds and uses the firm's weights, so each of the firm's portfolios becomes a draw of its own size from its channel's pool. The within-industry benchmark reassigns the same technologies across channels and owners. Each benchmark has 499 draws. For each distance and benchmark, I count the firm-years with a distance above its benchmark distance and compare that share with the same share in each draw, year by year. Each combination of distance and benchmark has one family of tests over the five years and the two weightings, and I apply Holm's adjustment (Holm, 1979) to each family's p-values. Table 4 summarizes the references.

The pattern in section 4.3 extends to the industry comparison. Firms keep a *technology mix of their own* within their industry, in their open-source code as in their patents. Under event weights, a firm's open-source code lies further from its industry's open-source code than the industry-pool benchmark places it, in 69% of firm-years (Holm-adjusted $p = 0.04$), and its patents lie further from its industry's patents than that benchmark places them, in 98.9% of firm-years ($p = 0.02$). The firm's own distance, by contrast, shows no clear difference from the industry-pool benchmark. That benchmark draws each of the firm's portfolios from its own channel's pool, so it preserves the industry's difference between open-source code and patents. The firm's own distance lies above that benchmark distance in 52.3% of firm-years ($p = 0.4$). Against the within-industry benchmark, which reassigns technologies across channels, the firm's open-source code and patents lie further apart in 78.7% of firm-years ($p = 0.04$), and the industry's own open-source code and patents do so in 99.8% of firm-years ($p = 0.04$). The pattern holds when the firm is compared with the industry's

leaders (Appendix F.2) or with all other owners (Appendix F.3).

As a firm adds items, each of its portfolios comes nearer its industry's portfolios, but more slowly than the industry-pool benchmark would bring it (Appendix Table G12). I estimate equation (1) for each of the five distances that involve the industry, with the three counts and firm and year fixed effects as the only regressors (Appendix H). As section 5.1 predicts, the benchmark distance falls faster because the randomization benchmark draws as many items as the firm holds. Within a firm and at the firm's own counts, each excess rises with the firm's count in the channel being compared with the industry. A doubling of the firm's open-source projects is associated with a change of 0.012 (standard error 0.005) in the excess from its open-source code to its industry's open-source code, and of 0.009 (standard error 0.002) in the excess to its industry's patents. A doubling of the firm's patents is associated with a change of 0.010 (standard error 0.003) in the excess from its patents to its industry's open-source code, and of 0.042 (standard error 0.003) in the excess to its industry's patents. None of the four excesses shows a clear association with the other channel's count. Across firms in the same industry-year, the excesses from the firm's open-source code are larger, and those from its patents are smaller, in larger programs, each at the firm's own counts (Appendix Table G11). As in section 5.2, these coefficients at the firm's own counts do not compare firms of different sizes.

5.6 Summary

This section asked whether the separation of section 4 reflects each firm's technology mix or the size of its open-source program. If open-source code were a direct substitute for a firm's own patents, it would lie only in technologies the firm patents. Instead, more than twice as much of it lies outside those technologies as under the randomization benchmark. Technology norms reproduce about half of the excess, and the remainder might reflect what firms choose to put into open-source code. Larger programs are measured with more items, so the benchmark distance falls with the count. The observed distance falls more slowly, so at each firm's own counts the excess rises with the count of open-source projects, as it does in simulated firms that keep one technology mix at every size, the second case in Table 6. At equal numbers of items, the excess does not rise within a firm and falls slightly across firms. The data are thus consistent with each firm keeping one mix at every size of its program. Across firms, the excess differs most clearly by industry. Size and research intensity show no clear association with it. Relative to later entrants, firms senior in the panel have open-source code further from their patents than the benchmark's draws produce. Relative to later starters, firms senior in open source have it nearer. Neither difference in the excess is clear with standard errors clustered by firm. Across a firm's repositories, those with more outside participant-years keep slightly less of their code outside the patent support than the benchmark gives, which is not consistent with the separation that Henkel et al. (2013) recommend. Within its industry, each of the firm's portfolios also lies further from the industry's portfolio in the same channel than the industry-pool benchmark places it, in most firm-years.

6 Conclusion

I link large firms to their open-source code and USPTO patents, and I use a language model to place the open-source code in the patent classification alongside the patents. Large firms' open-source code is concentrated in software engineering and computing, while their patents cover computing, communications and the physical sciences. I ask whether firms reveal through open-source code the technologies they protect with patents, or a separate set. The comparison shows that they also reveal technologies they do not patent. In 97% of firm-years, a firm's open-source code lies further from its patents in technology than the randomization benchmark built from the firm's own items places it. The excess over the randomization benchmark averages 23% of the firm's open-source code. This is the additional share that would have to move to other main groups to match the firm's patents, beyond what the randomization benchmark implies. Technology norms account for about half of the excess, 49%, and most of the rest, 50%, is the remainder, which might reflect what each firm chooses to open, technology by technology (section 4.4).

Firms that publish more open-source code have it nearer their patents, but the randomization benchmark at their own counts brings it nearer faster. None of 1,000 further draws of the randomization benchmark built from the firms' own technologies produces a slope as flat as the observed one (section 5.2). The excess therefore rises with the count of open-source projects at each firm's own counts. Simulated firms that keep a technology mix of their own at every size show the same rise, so the data are consistent with each firm keeping such a mix at every size of its open-source program. Across firms, industry is associated with the excess, while size and R&D intensity show no clear association with it (section 5.3). Within a firm, the distances from each of its portfolios to its industry's portfolios fall as that portfolio grows, but more slowly than the benchmark distances from the industry's pool. The benchmark distance falls faster because of the small-unit bias (section 5.5). If firms used open-source code as a direct substitute for their own patents, it would lie only in technologies they patent. Instead, the share of their open-source code outside those technologies is more than twice the corresponding share under the randomization benchmark.

The measure has limits. Distances depend on portfolio counts, so I compare each observed distance with its benchmark distance. The labels come from a language model, and checks against other classifications find agreement in many cases, but not all (section 3.5). The data omit the open-source code of accounts I cannot link to a firm. Despite these limits, the results show a consistent pattern. Future work can quantify how much these limits affect the pattern and assess ways to address them.

One question comes next. Patent examination for software became easier in two steps (Toole and Pairolero, 2020b, p. 14), a small one at the USPTO's April 2018 memorandum (U.S. Patent and Trademark Office, 2018) and a larger one at its January 2019 guidance on patent eligibility (U.S. Patent and Trademark Office, 2019, p. 50). The guidance binds examiners rather than courts (U.S. Patent and Trademark Office, 2019, p. 51), so the change concerns examination. Does easier patent examination for software change which technologies firms release as open-source code? Exposure to the guidance differs across technologies. A choice that varies across a firm's technologies leaves

a remainder, and a choice blind to technology leaves none (section 4.4). An incentive that shifts unevenly across technologies can therefore help tell the reasons for opening apart. Two checks would come before any answer. One is whether easier examination reached these firms' own applications and grants. The other is whether the firms' yearly releases in the more exposed technologies already moved in the years before 2018.

The data allow further uses. They can show how adding open-source code to a firm's technological position changes the spillover and rivalry pools of Jaffe (1986) and Bloom et al. (2013) (Appendix G), and who contributes to firms' open-source projects. Future work can test how far labeling both channels with one language model, hand-coded labels and dated snapshots of open-source code improve the measure.

One classification for open-source code and patents gives a fuller record of what firms know and what they share.

Use of AI tools

I used language models for three data-construction tasks and report the checks or bounds for each. I built a search agent with Claude Code using GLM-5.3-Flash and DuckDuckGo's web search to find firms' GitHub accounts. I accepted accounts using fixed criteria. A recall check bounds the share of firms whose accounts the search missed (section 3.2). In the patent crosswalk, a language model assisted the identification of the firm for 6.91% of all firm–patent links. I required evidence for the exact identifiers in each link (section 3.2). I used GLM-5.3-Flash to label the open-source projects. In section 3.5, I check the labels against the EPO's text categorizer and, by running DeepSeek-V4-Flash on patents, against the examiners' labels. The AI assistants Claude Code and Codex also helped with coding, editing and drafting. I reviewed all text and results and take responsibility for them.

References

Abbasiharofteh, M., Castaldi, C., and Petralia, S. (2026). A concordance between patent and trademark classes to link technologies to markets. *Scientific Data*, 13:1190.

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data. Technical Paper TP-2002-06, U.S. Census Bureau, Longitudinal Employer-Household Dynamics Program.

Aksoy-Yurdagul, D. (2025). Generating value by working with user communities: An analysis of financial market returns to corporate open source code contributions. *R&D Management*.

Alexy, O., George, G., and Salter, A. J. (2013). Cui bono? The selective revealing of knowledge and its implications for innovative activity. *Academy of Management Review*, 38(2):270–291.

Allen, R., Burgess, S., Davidson, R., and Windmeijer, F. (2015). More reliable inference for the dissimilarity index of segregation. *Econometrics Journal*, 18(1):40–66.

Alstott, J., Triulzi, G., Yan, B., and Luo, J. (2017). Mapping technology space by normalizing patent networks. *Scientometrics*, 110(1):443–479.

An, M., Kim, S., Sung, S., and Yoo, D. (2026). A context-aware framework for strategic patent intelligence: Keyword expansion and non-patent literature classification. *Online Information Review*, 50(5):962–980.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.

Argente, D., Baslandze, S., Hanley, D., and Moreira, S. (2025). Patents to products: Product innovation and firm dynamics. Working Paper 34592, National Bureau of Economic Research.

Arora, A., Belenzon, S., and Sheer, L. (2021). Matching patents to Compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures. *Research Policy*, 50(5):104217.

Arts, S., Cassiman, B., and Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Arts, S., Cassiman, B., and Hou, J. (2023). Position and differentiation of firms in technology space. *Management Science*, 69(12):7253–7265.

Autor, D., Chin, C., Salomons, A., and Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. *Quarterly Journal of Economics*, 139(3):1399–1465.

Baker, S. and Mezzetti, C. (2005). Disclosure as a strategy in the patent race. *Journal of Law and Economics*, 48(1):173–194.

Balla, S., Zacchiroli, S., Degueule, T., Falleri, J.-R., and Robbes, R. (2026). DRAGON: Robust classification for very large collections of software repositories. arXiv:2602.09071, version 1.

Bar, T. (2006). Defensive publications in an R&D race. *Journal of Economics & Management Strategy*, 15(1):229–254.

Bar, T. and Leiponen, A. (2012). A measure of technological distance. *Economics Letters*, 116(3):457–459.

Bar-Gill, O. and Parchomovsky, G. (2003). The value of giving away secrets. *Virginia Law Review*, 89(8):1857–1895.

Barkemeyer, D., Bena, J., and Li, X. (2026). Technology life cycles: Investment, growth, and firm value. Working paper 6350538, SSRN.

Baron, J. and Pohlmann, T. (2018). Mapping standards to patents using declarations of standard-essential patents. *Journal of Economics & Management Strategy*, 27(3):504–534.

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025). Inference for regression with variables generated by AI or machine learning. Technical report, arXiv:2402.15585.

Bello, P., Casarico, A., and Nozza, D. (2026). Research similarity and women in academia. *Economic Journal*, 136(679):2339–2368.

Bena, J., Ferreira, M. A., Matos, P., and Pires, P. (2017). Are foreign investors locusts? The long-term effects of foreign institutional ownership. *Journal of Financial Economics*, 126(1):122–146.

Benner, M. and Waldfogel, J. (2008). Close to you? Bias and precision in patent-based measures of technological proximity. *Research Policy*, 37(9):1556–1567.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6):3253–3285.

Brea, E. (2024). The yin yang of AI: Exploring how commercial and non-commercial orientations shape machine learning innovation. *Research Policy*, 53(6):105008.

Breschi, S., Lissoni, F., and Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1):69–87.

Cabral, L. M. B. (2002). Increasing dominance with no efficiency effect. *Journal of Economic Theory*, 102(2):471–479.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Cao, Y., Mkrtchyan, A., Tucker, J. W., and Wan, C. (2026). Corporate disclosure of innovation on GitHub. SSRN Working Paper 5932514, Donald G. Costello College of Business at George Mason University. First draft June 2025; this draft March 2026.

Carlson, N. A. and Burbano, V. (2026). The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*, 47(3):699–725.

Carrington, W. J. and Troske, K. R. (1997). On measuring segregation in samples with small units. *Journal of Business & Economic Statistics*, 15(4):402–409.

Chen, W., Jin, F., and Xue, L. (2022). Flourish or perish? The impact of technological acquisitions on contributions to open-source software. *Information Systems Research*, 33(3):867–886.

Chen, Y. (2025). Open or closed? Source-sharing strategies in AI competition. SSRN Working Paper 5830922; working paper, November 2025, University of Colorado Boulder.

Cockburn, I. M. and MacGarvie, M. J. (2011). Entry and patenting in the software industry. *Management Science*, 57(5):915–933.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not). Working Paper 7552, National Bureau of Economic Research.

Coleman, B., Fronk, K., and Valentine, K. (2025). Corporate adoption of an open innovation strategy: Evidence from GitHub. SSRN Working Paper 5068865, June 2025 version.

Comino, S. and Manenti, F. M. (2022). Patent portfolios and firms' technological choices. *Journal of Economics*, 137(2):97–120.

Conti, A., Peukert, C., and Roche, M. P. (2025). Beefing IT up for your investor? Engagement with open source communities, innovation, and startup funding: Evidence from GitHub. *Organization Science*, 36(4):1551–1573.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dell, M. (2025). Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58.

Dell, M. and Rambachan, A. (2026). The measurement revolution? credible measurement and inference in the age of AI. arXiv:2608.23524; NBER Working Paper 35744.

D'Haultfœuille, X. and Rathelot, R. (2017). Measuring segregation on small units: A partial identification analysis. *Quantitative Economics*, 8(1):39–73.

Dorner, M. and Harhoff, D. (2018). A novel technology-industry concordance table based on linked inventor-establishment data. *Research Policy*, 47(4):768–781.

Duan, J. and Pelger, M. (2026). Inference with AI-generated covariates. Working Paper 35481, National Bureau of Economic Research.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217.

Duranton, G. and Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4):1077–1106.

Economides, N. and Katsamakas, E. (2006). Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry. *Management Science*, 52(7):1057–1071.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Egami, N., Hinck, M., Stewart, B., and Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems 36*, pages 68589–68601.

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5):889–927.

Emer, L., Lippi, M., Mina, A., and Vandin, A. (2026). Large language models, encoder architectures and hybrid approaches for patent classification. arXiv:2601.23200, version 2.

Emery, L. P., Lim, C., and Ye, S. (2025). The private value of open-source innovation. SSRN Working Paper 5054880, revised November 2025.

Fosfuri, A., Giarratana, M. S., and Luzzi, A. (2008). The penguin has entered the building: The commercialization of open source software products. *Organization Science*, 19(2):292–305.

Gambardella, A. and Torrisi, S. (1998). Does technological convergence imply convergence in markets? Evidence from the electronics industry. *Research Policy*, 27(5):445–463.

Ganglmair, B. and Kann, A. (2025a). Looking for innovation beyond the patent system: Evidence from research disclosures. Slides presented at the DRUID 2025 Conference, June 27, 2025.

Ganglmair, B. and Kann, A. (2025b). Looking for innovation beyond the patent system: Evidence from research disclosures. Discussion Paper 656/2025, CRC TR 224, University of Bonn and University of Mannheim.

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Gill, D. (2008). Strategic disclosure of intermediate research results. *Journal of Economics & Management Strategy*, 17(3):733–758.

Goldschlag, N., Lybbert, T. J., and Zolas, N. J. (2020). Tracking the technological composition of industries with algorithmic patent concordances. *Economics of Innovation and New Technology*, 29(6):582–602.

Gong, X. (2026). Firms, patents and open source code: Two crosswalks. Working paper.

Gortmaker, J. (2025). Open source software policy in industry equilibrium. Working paper, January 19, 2025; jeffgortmaker.com.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Greenstein, S. and Nagle, F. (2014). Digital dark matter and the economic contribution of Apache. *Research Policy*, 43(4):623–631.

Habibi, M. (2025). Open sourcing GPTs: Economics of open sourcing advanced AI models. arXiv:2501.11581.

Haese, J. and Peukert, C. (2026). Open at the core: Moving from proprietary technology to building a product on open source software. *Management Science*, 72(5):4173–4199.

Hall, B. H. (2005). A note on the bias in Herfindahl-type measures based on count data. *Revue d'Économie Industrielle*, 110(1):149–156.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., and Taska, B. (2023). Remote work across jobs, companies, and space. NBER Working Paper 31007, National Bureau of Economic Research. Revised January 2026.

Harhoff, D., Henkel, J., and von Hippel, E. (2003). Profiting from voluntary information spillovers: How users benefit by freely revealing their innovations. *Research Policy*, 32(10):1753–1769.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *TEST*, 27(4):811–825.

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, 35(7):953–969.

Henkel, J., Baldwin, C. Y., and Shih, W. (2013). IP modularity: Profiting from innovation by aligning product architecture with intellectual property. *California Management Review*, 55(4):65–82.

Henkel, J., Schöberl, S., and Alexy, O. (2014). The emergence of openness: How and why firms adopt selective revealing in open innovation. *Research Policy*, 43(5):879–890.

Hoberg, G. and Phillips, G. M. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hoffmann, M., Nagle, F., and Zhou, Y. (2024). The value of open source software. Working Paper 24-038, Harvard Business School.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horstmann, I., MacDonald, G. M., and Slivinski, A. (1985). Patents as information transfer mechanisms: To patent or (maybe) not to patent. *Journal of Political Economy*, 93(5):837–858.

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76(5):984–1001.

Jaffe, A. B. (1989). Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2):87–97.

Johnson, J. P. (2014). Defensive publishing by a leading firm. *Information Economics and Policy*, 28:15–27.

Kamien, M. I. and Zang, I. (2000). Meet me halfway: Research joint ventures and absorptive capacity. *International Journal of Industrial Organization*, 18(7):995–1012.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2):665–712.

Korkmaz, G., Santiago Calderón, J. B., Kramer, B. L., Guci, L., and Robbins, C. A. (2024). From GitHub to GDP: A framework for measuring open source software innovation. *Research Policy*, 53(3):104954.

Landsman, W. R., Lin, C., Rabetti, D., and Zhang, C. (2026). Beyond coding: Open-source innovation disclosure and startup growth. SSRN Working Paper 5840682.

Lee, J.-S. and Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. *World Patent Information*, 61:101965.

Lerner, J. (1995). Patenting in the shadow of competitors. *Journal of Law and Economics*, 38(2):463–495.

Lerner, J. and Schankerman, M. (2010). *The Comingled Code: Open Source and Economic Development*. MIT Press, Cambridge, MA.

Lerner, J. and Tirole, J. (2002). Some simple economics of open source. *Journal of Industrial Economics*, 50(2):197–234.

Lerner, J. and Tirole, J. (2005). The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives*, 19(2):99–120.

Leten, B., Belderbos, R., and Van Looy, B. (2016). Entry and technological performance in new technology domains: Technological opportunities, technology competition and technological relatedness. *Journal of Management Studies*, 53(8):1257–1291.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078.

Lin, Y.-K. and Rai, A. (2024). The scope of software patent protection in the digital age: Evidence from Alice. *Information Systems Research*, 35(2):657–672.

Llanes, G. (2019). Competitive strategy for open and user innovation. *Journal of Economics & Management Strategy*, 28(2):280–297.

Lu, Q., Myers, A., and Beliveau, S. (2017). USPTO patent prosecution research data: Unlocking office action traits. Economic Working Paper 2017-10, U.S. Patent and Trademark Office, Office of the Chief Economist.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2026). Large language models: An applied econometric framework. *Annual Review of Economics*, 18:283–316.

Lybbert, T. J. and Zolas, N. J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, 43(3):530–542.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

Mazza, A. and Punzo, A. (2015). On the upward bias of the dissimilarity index and its corrections. *Sociological Methods & Research*, 44(1):80–107.

Mei, K. (2025). Why do firms and workers engage in open-source software development? Job market paper, McCombs School of Business, University of Texas at Austin; SSRN Working Paper 5698823.

Nagle, F. (2018). Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. *Organization Science*, 29(4):569–587.

Nagle, F. (2019). Open source software and firm productivity. *Management Science*, 65(3):1191–1215.

Nanba, H. and Takezawa, T. (2009). Classification of research papers into a patent classification system using two translation models. In Kan, M.-Y. and Teufel, S., editors, *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 27–35, Suntec City, Singapore. Association for Computational Linguistics.

Osborne, C., Daneshyan, F., He, R., Ye, H., Zhang, Y., and Zhou, M. (2025). Characterising open source co-opetition in company-hosted open source software projects: The cases of PyTorch, TensorFlow, and Transformers. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–30.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632.

Papke, L. E. and Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, 145(1–2):121–133.

Parchomovsky, G. (2000). Publish or perish. *Michigan Law Review*, 98(4):926–952.

Parker, G. and Van Alstyne, M. (2018). Innovation, openness, and platform control. *Management Science*, 64(7):3015–3032.

Patel, P. and Pavitt, K. (1995). Technological competencies in the world's largest firms: Characteristics, constraints and scope for managerial choice. IIASA Working Paper WP-95-066, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Petralia, S. (2024). GitPat: A database linking open source contributions & patenting activity of organizations. Papers in Evolutionary Economic Geography 2437, Utrecht University. Revised November 2024. http://econ.geo.uu.nl/peeg/peeg2437.pdf.

Petralia, S. (2025). Open source software as digital platforms to innovate. *Journal of Economic Behavior and Organization*, 237:107109.

Petralia, S. and Boschma, R. (2026). Actionable disclosure in patents: How open code reduces spatial barriers to knowledge diffusion. Papers in Evolutionary Economic Geography 26.11, Utrecht University. SSRN Working Paper 7090442, version of 9 July 2026.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39.

Podolny, J. M., Stuart, T. E., and Hannan, M. T. (1996). Networks, knowledge, and niches: Competition in the worldwide semiconductor industry, 1984–1991. *American Journal of Sociology*, 102(3):659–689.

Qiao, M. and Huang, K.-W. (2025). Correcting measurement error in regression models with variables constructed from aggregated output of data mining models. *MIS Quarterly*, 49(1):29–60.

Raymond, E. S. (2001). *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly & Associates, Sebastopol, CA, revised edition. The essays are also at http://www.catb.org/~esr/writings/cathedral-bazaar/.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Santos Silva, J. M. C. and Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, 88(4):641–658.

Singh, S. P., Ramasubbu, N., Kemerer, C. F., and Hydari, M. Z. (2026). Profitability of open-source software product development. *Information Systems Research*. Forthcoming; cited in the August 2026 version.

Sørensen, J. B. and Stuart, T. E. (2000). Aging, obsolescence, and organizational innovation. *Administrative Science Quarterly*, 45(1):81–112.

Spinellis, D., Kotti, Z., Kravvaritis, K., Theodorou, G., and Louridas, P. (2020). A dataset of enterprise-driven open source software. In *Proceedings of the 17th International Conference on Mining Software Repositories*, pages 533–537.

Stuart, T. E. and Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1):21–38.

Szczygielski, K., Lewkowicz, J., and Płoszaj, A. (2025). Thematic coherence of innovation systems: A patent-based index. *Economics of Innovation and New Technology*, 35(3):424–441. Published online 8 April 2025; issue dated 2026.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):285–305.

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior and Organization*, 23(1):1–30.

Toole, A. A. and Pairolero, N. A. (2020a). Adjusting to Alice: USPTO patent examination outcomes after Alice Corp. v. CLS Bank International. IP Data Highlights 3, U.S. Patent and Trademark Office, Office of the Chief Economist.

Toole, A. A. and Pairolero, N. A. (2020b). Supplementary material: Adjusting to Alice: USPTO patent examination outcomes after Alice Corp. v. CLS Bank International. IP Data Highlights 3, Supplementary Material, U.S. Patent and Trademark Office, Office of the Chief Economist. April 2020.

U.S. Patent and Trademark Office (2018). Changes in examination procedure pertaining to subject matter eligibility, recent subject matter eligibility decision (Berkheimer v. HP, Inc.). Memorandum, U.S. Patent and Trademark Office. From the Deputy Commissioner for Patent Examination Policy to the Patent Examining Corps, 19 April 2018. https://www.uspto.gov/sites/default/files/documents/memo-berkheimer-20180419.PDF; accessed 3 October 2026.

U.S. Patent and Trademark Office (2019). 2019 revised patent subject matter eligibility guidance. Federal Register notice 84 Fed. Reg. 50–57, U.S. Department of Commerce, Patent and Trademark Office. Published and effective 7 January 2019, document 2018-28282. https://www.federalregister.gov/documents/2019/01/07/2018-28282/2019-revised-patent-subject-matter-eligibility-guidance; accessed 3 October 2026.

Valentine, K., Zhang, J. L., and Zheng, Y. (2025). Strategic scientific disclosure: Evidence from the Leahy–Smith America Invents Act. *Journal of Accounting Research*, 63(4):1723–1755.

vom Stein, N., Sick, N., and Leker, J. (2015). How to measure technological distance in collaborations—the case of electric mobility. *Technological Forecasting and Social Change*, 97:154–167.

von Hippel, E. and von Krogh, G. (2003). Open source software and the "private-collective" innovation model: Issues for Organization Science. *Organization Science*, 14(2):209–223.

Wang, S., McCormick, T. H., and Leek, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275.

Webb, M. (2020). The impact of artificial intelligence on the labor market. Working paper, Stanford University. January 2020 version; SSRN Working Paper 3482150.

Wen, W., Forman, C., and Graham, S. J. H. (2013). The impact of intellectual property rights enforcement on open source software project success. *Information Systems Research*, 24(4):1131–1146.

West, J. (2003). How open is open enough? melding proprietary and open source platform strategies. *Research Policy*, 32(7):1259–1285.

Won, P. Y. (2026a). Commoditize your complement: Strategic open-source releases and the cannibalization of rival firms. Conference draft, May 2026. https://www.paris-december.eu/sites/default/files//papers/2026/7456_sutari4697_2026_complete.pdf.

Won, P. Y. (2026b). Commoditize your complement: Strategic open-source releases and the cannibalization of rival firms. SSRN Working Paper 6502879, version of June 2026. The SSRN record now serves the September 2026 revision; the June version's Proposition 1 appears in the same words in the May 2026 conference draft.

Won, P. Y. (2026c). Commoditize your complement: Who bears the cost of free software? SSRN Working Paper 6502879, revision of September 2026.

Xu, K. and Quispe, A. (2026). Industry classification of GitHub repositories using the North American Industry Classification System (NAICS). arXiv:2607.06505.

Yang, W. (2026). The financial implications of open innovation: Evidence from stock market reactions to public companies' open source software releases. *Research Policy*, 55(7):105500.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.

Zhao, Y. (2025). Open-source generative AI and firm value: Evidence from the release of DeepSeek. https://afajof.org/management/viewp.php?n=200088. Working paper, July 2025, Singapore Management University.

Ziedonis, R. H. (2004). Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. *Management Science*, 50(6):804–820.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A The Bar–Leiponen distance: derivations

This appendix gives the derivations behind the paper's distance argument. The paper's measure is the distance of Bar and Leiponen (2012), defined in section 3; every statement below is about it or derived from it. Lemma 4 writes the same distance in the open-source share within each technology; that form is an intermediate, used for the formula of technology norms alone (section 4.4) and in the proof of Proposition 3. The main text states the direction proposition, Proposition 3, in one sentence and unit breadth, Proposition 4, in words (section 3.4); both are stated and proved in full here (Appendices A.6 and A.7), as are the bounds from the GitHub classes and a distance to another portfolio (Appendices A.11 and A.12). The last three subsections give the six distances among the firm's two portfolios and its industry's, the references they are read against, and the floors behind the readings against the firm's own objects (Appendices A.13, A.14 and A.15); the results are in Appendix F. Every statement is an identity about shares or a consequence of moving one share. Nothing here is a statement about why any firm acts.

Throughout, a firm and a year are fixed, and the subscripts i and t are dropped where nothing is lost; restated results keep the main text's notation. For a vector v over technologies and a set S of them, $v(S) = \sum_{c \in S} v_c$, and $x^+ = \max\{x, 0\}$.

A.1 Definition and three equivalent forms

Let $p = (p_c)$ and $o = (o_c)$ be the firm's patent and open-source share vectors over the CPC main groups c, built as in section 3: over the five-year window ending in the year, with repositories weighted by their events. Each lies on the simplex: $p_c, o_c \geq 0$ and $\sum_c p_c = \sum_c o_c = 1$.

Definition 1 (The Bar–Leiponen distance)**.** $D = 1 - \sum_c \min\{p_c, o_c\}$.

This is the min-complement distance of Bar and Leiponen (2012, Definition 1, p. 457), with a firm's two channels in place of two firms' patent portfolios. In two cross-industry collaborations it differs little from Euclidean measures at the IPC section level and ranks pairs differently at the class level; vom Stein et al. (2015, pp. 165–166) recommend it for cross-industry settings while stating that its superiority has yet to be confirmed. The algebra below does not use how the shares were built, so it holds as well for the equal-weight companion D^A, which counts each repository once per window.

Lemma 1 (Three forms)**.** *For any p and o on the simplex,*

$$D = \tfrac{1}{2} \sum_c |p_c - o_c| = \sum_c (o_c - p_c)^+ = \max_S \left[o(S) - p(S) \right],$$

and the maximum is attained at $S^+ = \{c : o_c > p_c\}$. By symmetry, $D = \sum_c (p_c - o_c)^+ = \max_S [p(S) - o(S)]$.

Proof. For real a and b, $\min\{a, b\} = \frac{1}{2}(a + b - |a - b|)$. Summing over c and using $\sum_c p_c = \sum_c o_c = 1$ gives $\sum_c \min\{p_c, o_c\} = 1 - \frac{1}{2} \sum_c |p_c - o_c|$, the first equality. Since $\sum_c (o_c - p_c) = 0$, the positive and

negative parts of $o - p$ have equal sums; together they make $\sum_c |o_c - p_c|$, so each is half of it, the second equality. For any set S, $o(S) - p(S) = \sum_{c \in S}(o_c - p_c) \le \sum_{c \in S}(o_c - p_c)^+ \le \sum_c (o_c - p_c)^+$, with equality throughout at $S = S^+$, the third. Exchanging the roles of p and o gives the symmetric forms. □

The first equality is Proposition 1 of Bar and Leiponen (2012, p. 458), whose proof (p. 459) splits the groups by which of the two shares is smaller. The third is the best-split form: the distance is the largest amount by which the firm's open-source activity exceeds its patent activity on any set of main groups. In the construction of Duncan and Duncan (1955, p. 211) it is the largest vertical distance between the segregation curve and the diagonal, reached at the units whose group proportion is at least the overall one; here those are the groups with $o_c > p_c$.

The displacement reading. Each group with $o_c > p_c$ holds $o_c - p_c$ more open-source share than the patent mix calls for, and each group with $o_c < p_c$ holds less. Moving exactly the excess out of the first set into the second matches the patent mix, and no smaller move does, since at least $(o_c - p_c)^+$ must leave every group. So D is the smallest share of the firm's open-source activity that would have to change main group for the two portfolios to coincide: the displacement reading of Duncan and Duncan (1955, p. 211), and the share of non-overlapping activity of Bar and Leiponen (2012, p. 458).

Lemma 2 (Bounds)**.** $0 \le D \le 1$. $D = 0$ *if and only if* $p = o$. $D = 1$ *if and only if no main group holds both channels, that is,* $p_c o_c = 0$ *for every* c.

Proof. Each $\min\{p_c, o_c\}$ lies between 0 and p_c, so $\sum_c \min\{p_c, o_c\} \in [0, 1]$. It equals $1 = \sum_c p_c$ exactly when $p_c \le o_c$ for every c, which, since both vectors sum to one, means $p = o$. It equals 0 exactly when every term is zero, that is, when in every group at least one of the two shares is zero. □

Remark 1 (A metric). On the simplex, D is a metric: it is nonnegative and zero only for identical vectors, it is symmetric, and it satisfies the triangle inequality. Bar and Leiponen (2012) derive these properties from their Proposition 1 (p. 458; proof, p. 459), since half the L_1 distance is a metric, and state the range $[0, 1]$ (p. 458).

A.2 The exact split into outside and inside parts

Let $\operatorname{supp} p = \{c : p_c > 0\}$ be the firm's observed patent support in the window.

Lemma 3 (Outside and inside)**.**

$$D = \underbrace{\sum_{c:\, p_c = 0} o_c}_{\text{OUT}} + \underbrace{\sum_{c:\, p_c > 0} (o_c - p_c)^+}_{\text{IN}}, \qquad \text{OUT} = 1 - o(\operatorname{supp} p).$$

Both parts lie in $[0, 1]$. $\text{OUT} = 0$ *exactly when all open-source activity lies inside the patent support, and* $\text{IN} = 0$ *exactly when* $o_c \le p_c$ *on the support.*

Proof. By Lemma 1, $D = \sum_c (o_c - p_c)^+$. Split the sum by whether $p_c = 0$: for such c, $(o_c - p_c)^+ = o_c$. Both parts are sums of nonnegative terms, and their total is $D \leq 1$. □

OUT is the open-source mass outside the firm's observed patent support; IN is the mismatch within it, the open-source share in patented groups beyond the patent share there. The identity holds firm-year by firm-year, so it holds for every weighted mean over firm-years. On the five-year window, over 3,291 firm-years, the mean distance, 0.811, is 0.379548 outside and 0.431782 inside, which sum, up to rounding, to its unrounded value, 0.81133. The outside part is thus 46.8% of the mean at main-group depth. At subgroup depth the mean is 0.873096, which is 0.534227 outside and 0.338869 inside, and the outside part is 0.6119 of it. The split organizes the later tests: the outside part is where the paper reads placement, as a property of the technology or of the firm and relative to technology neighbors and complementors (sections B.6 and E.8), and the inside part is where it reads co-location and order (the presence association of Appendix B.5, and Appendix E.8). The excess over the randomization benchmark sits mostly in the outside part: 63% of it under equal weights and 85.7% under event weights (section A.10). Three "outside" numbers are different objects. The 46.8% is the outside part's share of the distance, a mean over firm-years; the 63% and 85.7% are its share of the excess over the randomization benchmark; and the 9.62% of section 3.4 is the share of the panel's open-source mass, pooled over firms, that lies in main groups where its firm holds no patent mass in the same window. The mean weighs every firm-year equally; the pooled share weighs each by its open-source mass.

Remark 2 (Irrelevant classes). Neither part depends on how the open-source share outside the patent support is spread over groups: OUT depends on it only through its total, and IN not at all. So $D(p, o) = D(p, o')$ whenever o and o' agree on $\operatorname{supp} p$. This is the independence of irrelevant patent classes of Bar and Leiponen (2012, Definition 2 and Proposition 2, p. 458), with the patent portfolio as the focal one, and Lemma 3 makes it visible. Exchanging the channels gives the symmetric split, the patent share outside the open-source support plus the mismatch within it; the paper reads the first, since its facts and tests ask where code sits relative to patents.

Remark 3 (Resolution). Suppose a finer classification nests in the coarser one, and the finer share vectors add up to the coarser ones. Then D and OUT at the finer depth are each at least their values at the coarser depth: merging groups can only cancel differences of opposite sign, since $|\sum_{g \in c}(o_g - p_g)| \leq \sum_{g \in c} |o_g - p_g|$, and every subgroup of a main group without patent share has none itself. IN has no fixed direction. The two depths above move this way; whether the paper's subgroup vectors add up exactly to its main-group vectors is not checked here. The dependence of the index on the units chosen is the one Duncan and Duncan (1955, p. 216) note, adding that smaller units do not remove it.

A.3 The distance in the open-source share

Section 4.1 reads the Bar–Leiponen distance as the spread of the firm's open-source share across its technologies. The lemma behind that reading follows, with its proof in full, its conditions and the

invariance it states.

Lemma 4 (The Bar–Leiponen distance in the open-source share)**.** *Let $O_{ic}, P_{ic} \geq 0$ be firm i's open-source and patent mass in technology c, with totals $O_i, P_i > 0$, shares $o_{ic} = O_{ic}/O_i$ and $p_{ic} = P_{ic}/P_i$, and Bar–Leiponen distance $D_i = \frac{1}{2}\sum_c |o_{ic} - p_{ic}|$. Let $\theta_{ic} = (O_{ic} + P_{ic})/(O_i + P_i)$ be technology c's share of the firm's combined activity, $\sigma_{ic} = O_{ic}/(O_{ic} + P_{ic})$ the open-source share within it, and $\bar{\sigma}_i = \sum_c \theta_{ic}\sigma_{ic}$. Then*

$$o_{ic} - p_{ic} = \frac{\theta_{ic}\,(\sigma_{ic} - \bar{\sigma}_i)}{\bar{\sigma}_i\,(1 - \bar{\sigma}_i)}, \qquad \textit{so} \qquad D_i = \frac{\sum_c \theta_{ic}\,|\sigma_{ic} - \bar{\sigma}_i|}{2\,\bar{\sigma}_i\,(1 - \bar{\sigma}_i)}.$$

The identity holds for any common rescaling of either channel's mass, which changes θ and σ but not D_i.

In the proof the firm index i is dropped. Let $O_c, P_c \geq 0$ be the firm's open-source and patent mass in technology c, with totals $O = \sum_c O_c > 0$ and $P = \sum_c P_c > 0$, so that $o_c = O_c/O$ and $p_c = P_c/P$. Call c held when $O_c + P_c > 0$, and on held technologies let $\theta_c = (O_c + P_c)/(O + P)$, $\sigma_c = O_c/(O_c + P_c)$ and $\bar{\sigma} = \sum_c \theta_c\sigma_c$. A technology that is not held has $o_c = p_c = 0$, contributes nothing to D, and is left out of every sum below.

Proof of Lemma 4. The average share. $\sum_c \theta_c\sigma_c = \sum_c O_c/(O + P) = O/(O + P)$, so $\bar{\sigma} = O/(O + P)$. Under $O, P > 0$ it lies strictly between 0 and 1, and $\bar{\sigma}(1 - \bar{\sigma}) = OP/(O + P)^2$.

One technology. By definition $O_c = \theta_c(O + P)\sigma_c$ and $P_c = \theta_c(O + P)(1 - \sigma_c)$. Hence

$$o_c - p_c = \theta_c(O+P)\Big[\frac{\sigma_c}{O} - \frac{1 - \sigma_c}{P}\Big] = \frac{\theta_c(O + P)\big[(O + P)\sigma_c - O\big]}{OP} = \frac{\theta_c(O + P)^2(\sigma_c - \bar{\sigma})}{OP} = \frac{\theta_c(\sigma_c - \bar{\sigma})}{\bar{\sigma}(1 - \bar{\sigma})}.$$

The distance. Half the sum of the absolute values over held technologies gives $D = \sum_c \theta_c|\sigma_c - \bar{\sigma}|/(2\bar{\sigma}(1 - \bar{\sigma}))$.

Rescaling. For $\kappa, \lambda > 0$, replace every O_c by κO_c and every P_c by λP_c. The shares o and p do not change, so neither does D. The weights and shares do change: σ_c becomes $\kappa O_c/(\kappa O_c + \lambda P_c)$, and θ_c and $\bar{\sigma}$ move with it. The three steps above hold for any positive masses, so the second form, evaluated at the rescaled θ, σ and $\bar{\sigma}$, equals the same D. Since the identity holds technology by technology, each technology's term is itself unchanged. □

The identity has been checked algebraically and numerically for any rescaling of either channel, and several independent derivations reach it. Two consequences are used below. By Lemma 1,

$$D = \frac{\sum_c \theta_c(\sigma_c - \bar{\sigma})^+}{\bar{\sigma}(1 - \bar{\sigma})}, \tag{2}$$

and, since $\theta_c > 0$ on held technologies, $o_c > p_c$ exactly when $\sigma_c > \bar{\sigma}$: the set S^+ of Lemma 1 is the set of technologies whose open-source share exceeds the firm's average.

Remark 4 (Duncan and Duncan's two forms). Both forms are the dissimilarity index of Duncan and Duncan (1955, p. 211), with technologies as the units and the two channels as the groups, open source in the role of their minority. The first is their half-sum of the absolute differences between the two groups' shares. The second is their "weighted mean deviation" of the units' group proportions from the overall proportion, divided by the overall mean deviation: $2\bar{\sigma}(1-\bar{\sigma})$ is the mean absolute deviation of a binary variable with mean $\bar{\sigma}$. Duncan and Duncan give these characterizations without proof, as elementary (p. 211); the steps above supply it for this setting. The paper's measure is the first form, as Bar and Leiponen (2012) define it. The second is an intermediate only, used because it shows the distance as a spread of the open-source share.

Remark 5 (What the invariance buys). A technology's counted output can differ by channel: a repository and a patent need not record the same amount of activity. Yields common to all of the firm's technologies, one per channel, multiply each channel's mass by a constant and, by the lemma, leave D unchanged. Yields that vary across technologies, even with a common ratio, reweight the technologies and can move D.

Direction and depth. The distance measures strength, not direction, so *away* and *near* are read against the firm's own average share, not against its patent core (Appendix A.6). Its second form is the dissimilarity index of Remark 4. At fixed weights, moving a technology's share away from the firm's average never lowers the distance, and where no technology sits exactly at the average, a small move toward it lowers the distance (Proposition 3). The same distance thus arises from a firm that opens where it patents and from one that opens everywhere but its patents, and since deviations of opposite sign cancel in averages across firms, a firm-level reading looks at the spread and the tails of deviations, not their mean. Main group is the finest depth with an external check: the EPO's categorizer returns main groups, and no independent reference exists for subgroup labels, so subgroup depth is a comparison layer only. On the same firm-years it raises the mean distance by 0.062, as finer groups must, since merging groups can only cancel differences (Remark 3); the gap is a property of the classification's depth, not evidence about placement.

What the outside part measures. One identity says how the share of code outside a set of technologies relates to the firm's split between the channels. Call any set H of the firm's technologies a *footprint* and H^c its complement. Let $q_{H^c} = \theta(H^c)$ be the share of the firm's combined patent and open-source mass outside the footprint, $\sigma_{H^c} = O(H^c)/(O(H^c)+P(H^c))$ and $\sigma_H = O(H)/(O(H)+P(H))$ the pooled open-source shares outside and inside it, and $o(H^c)$ the share of the firm's open-source mass outside it, in the notation of Lemma 4.

Lemma 5 (The part outside a footprint). *(i)* $o(H^c) = q_{H^c}\sigma_{H^c}/\bar{\sigma}$, *and*

$$o(H^c) - q_{H^c} = \frac{q_{H^c}(1-q_{H^c})(\sigma_{H^c}-\sigma_H)}{\bar{\sigma}}.$$

(ii) If the footprint contains the firm's patent technologies, $\{c : p_c > 0\} \subseteq H$, *then* $D \geq o(H^c)$; *when*

it is that set, $o(H^c)$ is the distance's outside part.

Proof. (i) $O(H^c) = q_{H^c}(O+P)\sigma_{H^c}$ and $O = \bar{\sigma}(O+P)$ give the first form; $\bar{\sigma} = q_{H^c}\sigma_{H^c} + (1-q_{H^c})\sigma_H$ gives $q_{H^c}\sigma_{H^c} - q_{H^c}\bar{\sigma} = q_{H^c}(1-q_{H^c})(\sigma_{H^c} - \sigma_H)$. (ii) Every technology outside such a footprint has $p_c = 0$, so $o(H^c)$ is at most the distance's outside part, which is at most D (Lemma 3). □

The shares in (i) must be pooled, each the open-source mass over the combined mass of the set; an unweighted mean of technology shares does not give the identity. Part (ii) needs the footprint to contain the support of the patent portfolio in the same distance: a footprint built from the firm's earlier patents need not, and the bound can then fail. With the same window's support as the footprint, all mass outside it is open source, $\sigma_{H^c} = 1$, so (i) says only that the outside part is at least the combined mass outside the support. With a footprint fixed before the window, a different object that the paper keeps apart from the distance's outside part, (i) says that the share of code outside the footprint exceeds the share of activity there exactly when the firm opens a larger share of its activity outside the footprint than inside it.

A.4 Technology norms alone, for the Bar–Leiponen distance

Proposition 1 states the formula of section 4.4 at its proved scope. It takes every holder of technology c to open the same share f_c of its public output there. The public channel masses are then proportional to $\theta_c f_c$ and $\theta_c(1-f_c)$, so the distance below is the Bar–Leiponen distance between the firm's public-output portfolios, before counts and kernels; sections A.7 and A.8 add those. The proposition compares configurations at given output weights: it does not let θ respond to the shares.

Proposition 1 (Separation without a firm term)**.** *Let every holder of technology c open the same share $f_c \in (0,1)$ of its public output there; take the output weights θ_{ic} as given; and let Y_i take the value f_c with probability θ_{ic}, with mean $\bar{\sigma}_i$. (Y_i is a device: it is the open-source share of the technology of one unit of output, drawn with the weights θ_{ic}.) Then:*

(a) $D_i = \mathrm{E}|Y_i - \bar{\sigma}_i| / (2\bar{\sigma}_i(1-\bar{\sigma}_i))$;

(b) $D_i = 0$ *exactly when f_c is constant on the firm's technologies, those with $\theta_{ic} > 0$, and $D_i > 0$ otherwise;*

(c) at a common mean, a mean-preserving spread of Y_i does not lower D_i.

Proof. (a) Since $f_c \in (0,1)$, both channel masses are positive and $m = \sum_c \theta_c f_c = \bar{\sigma} \in (0,1)$. Lemma 4 gives $D = \sum_c \theta_c |f_c - \bar{\sigma}| / (2\bar{\sigma}(1-\bar{\sigma}))$, which is $\mathrm{E}|Y-m|/(2m(1-m))$, since Y takes the value f_c with probability θ_c.

(b) The denominator is positive, and $\mathrm{E}|Y-m| = 0$ exactly when $Y = m$ with probability one, that is, when f_c is the same on every technology with $\theta_c > 0$.

(c) Let Y' be a mean-preserving spread of Y: Y' has the law of $Y + Z$ with $\mathrm{E}[Z \mid Y] = 0$. Then $\mathrm{E}Y' = m$, so the denominator is unchanged, and by Jensen's inequality for the convex function $y \mapsto |y-m|$, conditionally on Y, $\mathrm{E}|Y'-m| = \mathrm{E}[\mathrm{E}[|Y+Z-m| \mid Y]] \geq \mathrm{E}|Y-m|$. □

Remark 6 (The scope: fixed output weights). The distance depends on the weights θ as well as on the shares. If the mix of what becomes public moves with the norms — for example, if a firm makes less of its output public in technologies whose norms change — a change in the norms moves the weights too, and the distance can fall as the norms spread. Part (c) is therefore a statement at fixed output weights.

Remark 7 (The fixed-portfolio world). Worlds in which every holder of a technology has the same share f_c are the case of technology norms alone (section 4.4). The fixed-portfolio world (section 4.2) is not of that kind: it fixes each firm's own whole-period pools, so it keeps whatever is specific to a firm and a technology, provided it does not change over time. Its reproduction of the excess's means shows that they need no change over time in what a firm holds and are compatible with fixed portfolios, and so are not evidence of a choice; it neither tests nor supports the formula of technology norms alone, whose own world is section 4.4's.

The element-by-element world. A technology-position world, in which each firm opens each technology it holds at that technology's common open-source share among the panel's other firms, redraws each unit's technologies element by element, so it does not use the randomization benchmark's device; section 4.4 reads the nested references, which do. With each firm's own units redrawn, it reproduces 58% of the excess under equal weights and 55.4% under event weights. With the redraw switched off, its closed form reproduces 27.7% under equal weights and lies below the randomization benchmark's reference under event weights (−19.8%). The closed form carries no sampling noise, while the data and the randomization benchmark do, and under any unbiased draw of the units the share is at least the closed form's, which bounds it from below. The world's distance exceeds the randomization benchmark's reference in 93.3% and 90.2% of firm-years and reaches the observed distance in 7.4% and 12.7%; the world and observed distances correlate across firm-years at 0.79 and 0.76, under equal and event weights (section 4.4). With the redraw of units switched off, the world's distance has a closed form, 0.5257 under both weights. The share of the excess it reproduces depends on how units are drawn: the world draws each element of a unit's profile independently, whereas the randomization benchmark moves whole observed vectors. In the world, the slope of the distance on the lagged repository count is −0.0346 under equal weights and −0.0379 under event weights, against −0.0204 and −0.0152 observed: technology position with the firm's own counts also produces part of the count slope. A main group that no other panel firm holds in the window has no common share; those cells are dropped and the weights renormalized, in 644 firm-years and 0.16% of a firm-year's weight on average.

A Poisson companion. The nested references of section 4.4 draw with the randomization benchmark's device. A companion fits each firm-year's counts by main group and channel by Poisson pseudo-likelihood (Santos Silva and Tenreyro, 2006, p. 645), on the same 2,472 firm-years. Every model keeps firm-year × technology and firm-year × channel effects, so each firm keeps its own technology vector and its own channel totals, as in the randomization benchmark. The

first has only those: within a firm-year the channel does not depend on technology, and its fitted distance is zero. The next two add, as offsets, the leave-one-firm-out technology norm and then the technology-by-industry norm that the nested references draw with. On counts normalized within each firm-year and channel, as in the mean distance, the technology norm's fitted portfolios reproduce 68.3% of the observed mean distance under equal weights and 64.8% under event weights, and the technology-by-industry norm's 69.4% and 65.8%. On the raw counts the technology norm removes 67.4% and 63.2% of the deviance that the firm-year effects alone leave, and the technology-by-industry norm's further share is negative, −9.8% and −10%: it fits the channel mix worse than the technology norm alone.

The companion's base differs from the nested references'. It starts at zero and reads shares of the observed mean distance; the nested references start at the randomization benchmark's mean and read shares of the excess over it, so the two are read side by side, not as one ratio. Both find that technology norms account for much of what they measure and that industry beyond technology adds little. What the companion leaves unreproduced, 42.4% and 46.8% of the deviance, includes count noise, which a Poisson deviance carries even under a correct model and the randomization benchmark's base absorbs, so it is not the firm's remainder. The companion has no intervals. Its check that the first model's fitted counts match each firm-year's margins failed, with largest relative errors of 0.002 and 0.0092 under equal and event weights. Because the three models separate by firm-year, they were also solved exactly, one firm-year at a time; that solution matches the fitted deviances and distances to rounding, and the shares above are its. No firm-cluster bootstrap was run.

A.5 Choice, characteristics and the threats to identification

Section 4.4 states this result in words.

Proposition 2 (What firm data identify)**.** *(i)* Within a firm. *If $a_i + s_c + u_{ic} = a'_i + s'_c + u'_{ic}$ for every technology c the firm holds, the laws of all of firm i's observed units coincide, and so do the laws of every statistic computed from them: its distance, and any reference that randomly assigns its own units, included.*

(ii) Across firms. *If the chosen component contains a part in the span of firm and technology effects, $\nu_{ic} = g_c + h_i + \nu^{\perp}_{ic}$ with $\nu^{\perp}$ orthogonal to that span, the recovered deviation is $\eta^{\perp} + \nu^{\perp}$: a choice made by every holder of technology c is absorbed into the norm s_c, and a choice applied to all of a firm's technologies into its openness a_i.*

Hence a departure relative to technology peers, and relative to the firm's own overall openness, can be identified; reading it as choice needs restrictions on hidden characteristics, on selection and on measurement.

Every reference that randomly assigns the firm's own units measures technology position and choice together, because part (i) is immediate: the shares depend on s_c and u_{ic} only through their sum. Part (ii) is exact for the projection. Part (i) is the version for the two channels of an observation

that D'Haultfœuille and Rathelot (2017, p. 39, fn. 2) make about segregation indices, and Ellison and Glaeser (1997, p. 897) about concentration indices: an index does not reveal the mechanism that produced it. Three threats bound every design built on the proposition: shared choice is absorbed into the technology, unchosen deviations look like choices, and observation artifacts correlated with the firm's mix, unit breadth among them, create apparent deviations; This subsection states them, with the remarks on what a firm's own panel can see.

This subsection gives the detail behind section 4.4: the two stories the decomposition holds, what a firm's own panel can see, and the three threats that bound every design built on Proposition 2.

Choice and characteristics. A firm's code can sit farther from its patents than a reference, or closer, for two kinds of reason. In the first story, *choice*, the firm places its code differently from what its technologies call for: *away*, farther from its patents than its technologies imply, or *near*, closer than they imply. In the second story, *characteristics*, the firm makes no special placement decision, but its distance still differs from other firms' distances, for one of four reasons:

- its technology position: technologies differ in how naturally they go to patents or to code;
- its volume: fewer units make the measured distance larger under random assignment;
- its kinds of units: a patent spans more technology groups than a repository, a repository that is active for years is one object and not several, and event weights let one busy repository dominate;
- its overall openness.

The firm's observed units are its patents and repositories, with their technology vectors, years and weights; a law draws them, and this law depends on the shares only through the two portfolios, so the firm's volume and kinds of units shape the measured distance. The counts carry choice too: the ratio of a firm's two channels contains a_i and u_{ic}.

What a firm's own panel can see. If a_i and s_c are fixed over time, the firm's own panel identifies changes in its deviations across its technologies, up to a firm-by-year constant: what such a panel can see is a change in the deviation. And if the suitability of a technology may differ by firm, no data identify the chosen part at all, since every chosen deviation then has an observationally equivalent suitability. From its own units, a firm with a chosen deviation is indistinguishable from a firm without one whose technology norms are shifted by that deviation.

The threats. Three threats bound every design built on the proposition. First, shared choice is technology (part (ii)): a rule common to a technology position can never be told apart from the position. Second, unchosen deviations look like choices: a firm that writes a different kind of code within the same group, faces a different patentability, has different staff or carries past choices forward deviates without choosing. A core defined from the firm's own earlier patents makes this worse, because persistent unchosen deviations then show up as a spurious *away*. Third, observation artifacts correlated with the firm's mix create apparent deviations: unit breadth, which separates

the portfolios even when choice ignores technology (Proposition 4), so that a test of deviations uses units placed in a single group or models the kernels; persistent repositories; and small counts in small technologies, which also inflate the spread of estimated deviations.

A.6 The direction proposition

The proof pools technologies into two groups. A two-group firm has combined weights t and $1-t$, with $0 < t < 1$, and open-source shares x and y; its average share is $m = tx + (1-t)y$, with $0 < m < 1$.

Lemma 6 (Two groups)**.** *The two-group firm's Bar–Leiponen distance is* $D_2(t, x, y) = t(1-t)|x - y|/(m(1-m))$. *For* $x > y$,

$$\frac{\partial D_2}{\partial x} = \frac{t(1-t)\big[y(1-y) + t^2(x-y)^2\big]}{(m(1-m))^2} > 0, \qquad \frac{\partial D_2}{\partial y} = -\frac{t(1-t)\big[x(1-x) + (1-t)^2(x-y)^2\big]}{(m(1-m))^2} < 0.$$

Proof. By Lemma 4 with two technologies, $o_1 - p_1 = t(x-m)/(m(1-m))$, and $x - m = (1-t)(x-y)$; with two groups the two differences cancel, so $D_2 = |o_1 - p_1|$. For $x > y$ write $g = m(1-m)$; since $\partial m/\partial x = t$, $\partial D_2/\partial x = t(1-t)\big[g - t(x-y)(1-2m)\big]/g^2$. With $d = t(x-y)$ and $m = y + d$, $g - d(1-2m) = (y+d)(1-y-d) - d(1-2y-2d) = y(1-y) + d^2$. Likewise $\partial m/\partial y = 1 - t$, and with $e = (1-t)(x-y)$ and $m = x - e$, $g + e(1-2m) = x(1-x) + e^2$, which gives the second derivative with its sign. Both brackets are positive because $x > y$, and $t(1-t) > 0$. □

Lemma 7 (The best split in the open-source share)**.** *For a set* S *of the firm's technologies with* $0 < \theta(S) < 1$, *let* $\sigma(S) = \sum_{c \in S} \theta_c \sigma_c / \theta(S)$ *be its pooled share. Then*

$$o(S) - p(S) = \frac{\theta(S)\big(\sigma(S) - \bar{\sigma}\big)}{\bar{\sigma}(1-\bar{\sigma})},$$

which is the signed distance of the two-group firm that pools S *and its complement. When* $D > 0$, $D = \max_S D_2\big(\theta(S), \sigma(S), \sigma(S^c)\big)$, *attained at* $S^* = \{c : \sigma_c > \bar{\sigma}\}$.

Proof. Summing Lemma 4 over S gives the displayed identity. Since $\bar{\sigma} = \theta(S)\sigma(S) + (1-\theta(S))\sigma(S^c)$, it is $t(x-m)/(m(1-m))$ with $t = \theta(S)$, $x = \sigma(S)$, $y = \sigma(S^c)$ and $m = \bar{\sigma}$: pooling leaves each group's channel masses unchanged, so this is the two-group firm's $o_1 - p_1$, and $|o(S) - p(S)| = D_2(\theta(S), \sigma(S), \sigma(S^c))$. By Lemma 1 and its symmetric form, $|o(S) - p(S)| \le D$ for every S, with equality at S^+, which is S^* by (2) and the remark after it. □

Proposition 3 (Direction)**.** *Hold the output weights* θ *fixed. For a firm with average open-source share* $\bar{\sigma}$*: if* $\sigma_j > \bar{\sigma}$, *raising* σ_j *does not lower the Bar–Leiponen distance; if* $\sigma_j < \bar{\sigma}$, *lowering* σ_j *does not lower it. Where no technology has* $\sigma_c = \bar{\sigma}$, *so that the split* S^* *is unique, a small move of* σ_j *toward* $\bar{\sigma}$ *lowers the Bar–Leiponen distance. In the portfolios, the first part holds for any change that adds open-source mass to one technology and removes patent mass from it, at fixed masses elsewhere (Remark 8).*

Proof of Proposition 3. Let $\sigma_j > \bar{\sigma}$. Then $j \in S^*$, and $D > 0$, so $0 < \theta(S^*) < 1$. Write $t = \theta(S^*)$, $y = \sigma(S^{*c})$ and $x = \sigma(S^*)$; then $x > \bar{\sigma} > y$. Raising σ_j at fixed θ raises x to some $x' > x$ and leaves t and y unchanged. After the move the distance is the maximum over all splits (Lemma 7), so it is at least $D_2(t, x', y)$; by Lemma 6, $D_2(t, x', y) \geq D_2(t, x, y)$; and $D_2(t, x, y)$ is the distance before the move, by Lemma 7. If $\sigma_j < \bar{\sigma}$, then $j \notin S^*$, and lowering σ_j lowers y at fixed t and x; the sign of $\partial D_2/\partial y$ gives the same chain.

For the converse, suppose no technology has $\sigma_c = \bar{\sigma}$. The average $\bar{\sigma}$ moves continuously with σ_j, so a small enough move leaves the sign of every $\sigma_c - \bar{\sigma}$ unchanged, hence S^* unchanged, and along the move $D = D_2(\theta(S^*), \sigma(S^*), \sigma(S^{*c}))$ by (2). A move of σ_j toward $\bar{\sigma}$ lowers x when $j \in S^*$ and raises y when $j \notin S^*$, and the strict signs of Lemma 6 make D fall. □

Remark 8 (The same result in the portfolios). By Lemma 1, $D = o(S^+) - p(S^+)$. Take $j \in S^+$ and a change that adds $\delta_O \geq 0$ to O_j and removes $\delta_P \in [0, P_j]$ from P_j, leaving every other technology's masses unchanged. Since $O(S^+) \leq O$, $o(S^+)$ becomes $(O(S^+) + \delta_O)/(O + \delta_O) \geq O(S^+)/O$; since $P(S^+) \leq P$ and $P - P_j > 0$ (as $p_j < o_j \leq 1$), $p(S^+)$ becomes $(P(S^+) - \delta_P)/(P - \delta_P) \leq P(S^+)/P$. The new distance is at least the new value of this split, so it is at least D. The case $j \notin S^+$ is the mirror. At fixed θ, raising σ_j is the case $\delta_O = \delta_P$. The argument needs the change to move the technology's two channel masses in opposite directions: a change that raises both can lower the distance.

Remark 9 (Direction relative to the firm's own average). *Away* and *near* are defined against the firm's own average share $\bar{\sigma}$, not against its patent core. A tilt toward the core raises the distance when the core was already the group with the higher open-source share, and lowers it otherwise. A direction for the distance is not a sign for a firm's deviation from its technology peers (section 4.4), which is measured against the technologies' norms rather than the firm's own average.

Remark 10 (Aggregation). A firm's excess over technology norms, its distance minus the distance it would have at its technologies' norms, reads as *away* when positive and *near* when negative. Excesses of opposite sign cancel in any average across firms, so an average excess of zero is consistent with every firm deviating, some one way and some the other, and does not mean that no firm chooses. A firm-level test should therefore look at the spread and the two tails of the deviations, not at their mean.

An example: openness alone has no sign. With two technologies of equal weight and open-source shares $\frac{1}{5}$ and $\frac{3}{5}$, the distance is $\frac{5}{12}$; raising the firm's openness to $\frac{1}{2}$ through the second technology, to shares $\frac{1}{5}$ and $\frac{4}{5}$, raises it to $\frac{3}{5}$, and raising it the same amount through the first, to $\frac{2}{5}$ and $\frac{3}{5}$, lowers it to $\frac{1}{5}$. The first move takes a share away from the firm's average and the second toward it, as Proposition 3 says (Appendix B.11).

A.7 Unit breadth

A unit whose technology is c records its mass over classification groups g through a row-stochastic kernel: $K_P(c,g)$ for a patent and $K_O(c,g)$ for a repository, with $\sum_g K(c,g) = 1$. Writing the technology masses as row vectors $(P_c)_c$ and $(O_c)_c$, the observed masses are $(P_c)_c K_P$ and $(O_c)_c K_O$, and, since a row-stochastic kernel preserves totals, the observed shares are pK_P and oK_O. Technology-blind choice means a common open-source share, $\sigma_c = \bar{\sigma}$ on every held technology ($u = 0$ with a constant share); then $p = o = \theta$.

Proposition 4 (Unit breadth)**.** *Suppose patents and repositories spread their mass over technology groups through kernels K_P and K_O, and choice is technology-blind: the firm's open-source share is the same in all its technologies ($u = 0$ with σ constant). Then the observed open-source share of group g is*

$$\sigma_g^{\text{obs}} = \frac{\bar{\sigma}\,[\theta K_O]_g}{(1-\bar{\sigma})\,[\theta K_P]_g + \bar{\sigma}\,[\theta K_O]_g}, \qquad \log\frac{\sigma_g^{\text{obs}}}{1-\sigma_g^{\text{obs}}} = \log\frac{\bar{\sigma}}{1-\bar{\sigma}} + \log\frac{[\theta K_O]_g}{[\theta K_P]_g},$$

where both brackets are positive, so the observed shares vary with the firm's own mix through the kernels; and the population Bar–Leiponen distance is $D^{\text{obs}} = \frac{1}{2}\sum_g |[\theta K_P]_g - [\theta K_O]_g|$, positive exactly when the kernels differ on the firm's mix, $\theta K_P \neq \theta K_O$. That the offset $\log([\theta K_O]_g/[\theta K_P]_g)$ is not in general in the span of firm and technology effects is shown by example; the general statement is qualitative.

Proof. With a common share, $P_c = (O+P)\theta_c(1-\bar{\sigma})$ and $O_c = (O+P)\theta_c\bar{\sigma}$ for the firm's total mass $O+P$, so the observed masses in group g are $(O+P)(1-\bar{\sigma})[\theta K_P]_g$ and $(O+P)\bar{\sigma}[\theta K_O]_g$; their ratio gives σ_g^{obs} and its log-odds. The observed shares are θK_P and θK_O, so Lemma 1 gives D^{obs}, and Lemma 2 gives its zero condition. □

The offset is additive in firm and group only in special cases; equal kernels make it zero. In an example with twelve technologies on a circle, patents spread over three groups and repositories over one, and no choice at all, the apparent deviations reach more than three units of log-odds and correlate with the firm's peak technology. "Differ" means differ on the firm's mix: kernels that differ but agree on it, $\theta K_P = \theta K_O$, give no distance. With each channel keeping weight w_X on its own group and spreading the rest by a common kernel M, the distance is $|w_P - w_O| \cdot \frac{1}{2}\|\theta - \theta M\|_1$, zero when $\theta M = \theta$. Beyond technology-blind choice, two bounds hold. A kernel common to both channels can only shrink the distance, $\frac{1}{2}\|(p-o)K\|_1 \le \frac{1}{2}\|p-o\|_1$, because K is row-stochastic. With different kernels the observed distance lies within D of the recording level $f = \frac{1}{2}\|pK_P - pK_O\|_1$, $f - D \le D^{\text{obs}} \le f + D$, by the triangle inequality through pK_O and the first bound. Recording is a level, not a floor: it can create separation, and it can hide it. Hence the main text's rule: a test of deviations uses units placed in a single group, or models the kernels.

A.8 Random assignment, counts and the reference

The paper's distance is computed from finitely many units, so it is a statistic of the shares those units are drawn from. Write $\hat{D}$ for the measured distance and D_{pop} for the distance between the population shares (D^* is kept for Carrington and Troske's level under random assignment, section A.9). The results below are about hard units, each placed in one group with equal weight, unless they say otherwise; the paper's units are not of that kind, which is why the paper reads its facts against a reference built from the firm's own units.

Random allocation. Under random allocation, with channel independent of technology, the firm's n_P patents and n_O repositories are independent draws from one mix θ, so $D_{\text{pop}} = 0$. Conditioning also on the firm's pooled technology counts gives the fixed-margin shuffle that Carrington and Troske (1997, pp. 403–405) name and simulate. Let $n_{\min} = \min\{n_P, n_O\}$ and $1/H = 1/n_P + 1/n_O$.

Lemma 8 (The level under random assignment is positive)**.** *Under random allocation,*

$$\mathrm{E}\hat{D} = B(n_P, n_O, \theta) = \tfrac{1}{2} \sum_c \sum_{x=0}^{n_P} \sum_{y=0}^{n_O} \Big| \frac{x}{n_P} - \frac{y}{n_O} \Big| \operatorname{Bin}(x; n_P, \theta_c) \operatorname{Bin}(y; n_O, \theta_c),$$

with

$$\sum_c \theta_c (1-\theta_c)^{n_{\min}} \le B \le \min\Big\{1,\ \frac{1}{2\sqrt{H}} \sum_c \sqrt{\theta_c(1-\theta_c)}\Big\},$$

and $B = 1 - \sum_c \theta_c^2$ when the smaller channel has one unit. So the expected distance is positive at every finite count unless the mix puts all its weight on one technology.

Proof. Let X_c and Y_c be the numbers of patents and repositories in technology c. By linearity of expectation each term of $\frac{1}{2}\sum_c |X_c/n_P - Y_c/n_O|$ depends only on its two independent binomial marginals, which gives B. For the lower bound, $\hat{D} = 1 - \sum_c \min\{X_c/n_P, Y_c/n_O\}$ and $\min\{X_c/n_P, Y_c/n_O\} \le (X_c/n_P)\mathbf{1}\{Y_c \ge 1\}$, whose expectation is $\theta_c(1-(1-\theta_c)^{n_O})$ by independence; summing, $B \ge \sum_c \theta_c(1-\theta_c)^{n_O}$, and by symmetry $B \ge \sum_c \theta_c(1-\theta_c)^{n_P}$, the larger of the two being the one at $n_{\min}$. It is positive as soon as some θ_c lies strictly between 0 and 1, that is, as soon as two technologies carry weight; with one technology $\hat{D} = 0$. For the upper bound, $X_c/n_P - Y_c/n_O$ has mean zero and variance $\theta_c(1-\theta_c)/H$, so its expected absolute value is at most $\sqrt{\theta_c(1-\theta_c)/H}$. With one repository, in technology c, $\hat{D} = 1 - X_c/n_P$, whose expectation averaged over c is $\sum_c \theta_c(1-\theta_c)$; one patent is symmetric. □

The expectation is not new: Allen et al. (2015, §2.1, p. 44) write the expected dissimilarity index as a sum over two independent multinomial allocations. The fixed-margin model, which conditions also on the firm's pooled technology counts, is the one Allen et al. (2015, p. 44) set aside and Carrington and Troske (1997, pp. 403–405) name and simulate; averaged over those counts, its mean is B. The upper bound falls as H grows at a fixed mix, so the level under random assignment vanishes only as both counts grow. The level is symmetric in the two counts, as the formula shows.

Lemma 9 (The measured distance is biased up)**.** *Whenever the measured shares are unbiased,* $\mathrm{E}\hat{p} = p$ *and* $\mathrm{E}\hat{o} = o$*,* $\mathrm{E}\hat{D} \geq D_{\mathrm{pop}}$*.*

Proof. $D(p, o) = \frac{1}{2}\|p - o\|_1$ is convex in the pair (p, o), being a norm of a linear map of it; Jensen's inequality gives $\mathrm{E}D(\hat{p}, \hat{o}) \geq D(\mathrm{E}\hat{p}, \mathrm{E}\hat{o})$. □

Gentzkow et al. (2019b, p. 1315) make the same point for any norm, and Mazza and Punzo (2015, Lemma 1, p. 86) for the dissimilarity index itself. For patent-based measures, Benner and Waldfogel (2008, NBER Working Paper 13322, abstract and p. 17) show that proximity computed from few patents is biased and imprecise, with sampling error overstating the distance between firms that are truly close, and Hall (2005, working paper of January 2005, pp. 1–3) that a Herfindahl index computed from few counts is biased, with a correction that depends only on the total count. The lemma needs only unbiased shares, so it covers fractional unit vectors, and weights independent of the groups, not only hard units. The level under random assignment itself is not a floor for $\mathrm{E}\hat{D}$ when $D_{\mathrm{pop}} > 0$: with unequal counts $\mathrm{E}\hat{D}$ can dip below it. The floor is D_{pop}.

How counts shape it. The following holds on stated models, not on the paper's firms. In exact toy calculations the level under random assignment rises with $1/n_P + 1/n_O$, so it is set mainly by the smaller count, though that sum does not pin it exactly. With equal-weight whole units it does not rise as either count grows; the records state this from an earlier review, without a new proof. With event weights it can rise: with patent and open-source shares both $(\frac{1}{2}, \frac{1}{2})$ and two equal-weight patents, adding a repository of weight 100 to two of weight 1 raises $\mathrm{E}\hat{D}$ from 3/8 to 101/204. Fractional patents spread over several groups, event weights and repositories that persist across years move the level under random assignment further, so no formula for hard units gives the paper's level under random assignment.

Why the reference is built from the firm's own units. Since the level under random assignment depends on the counts, the mix and the kinds of units, all of which differ across firms, a reference must hold them at the firm's own values. The randomization benchmark does so: it reallocates whole vectors among a firm's objects that start in the same year, keeping each object's channel, years and weights (section 4.2). It is the fixed-size redraw of Carrington and Troske (1997, §3.1, p. 405), generalized to fractional, weighted objects within strata.

Lemma 10 (A within-firm shuffle is centered under exchangeability)**.** *Let* V *assign vectors to the firm's objects, let* $\mathcal{G}$ *be the group of permutations of vectors among objects within the same stratum, and let* $R(V) = |\mathcal{G}|^{-1} \sum_{g \in \mathcal{G}} D(gV)$ *be the randomization benchmark's mean, with* D *computed from each object's fixed channel and weights. If the law of* V *is invariant under* $\mathcal{G}$*, that is, if object vectors are exchangeable across the firm's channels within strata, then* $\mathrm{E}[D(V) - R(V)] = 0$*; and, for a law with finitely many values,* $\mathrm{E}[D(V) \mid V \in \mathcal{G} \cdot v] = R(v)$ *for every orbit* $\mathcal{G} \cdot v$*.*

Proof. Invariance gives $\mathrm{E}\, D(gV) = \mathrm{E}\, D(V)$ for every $g \in \mathcal{G}$; averaging over g gives $\mathrm{E}\, R(V) = \mathrm{E}\, D(V)$. For a law with finitely many values, invariance gives $\Pr(V = v) = \Pr(V = gv)$ for every $g \in \mathcal{G}$,

so given its orbit V is uniform on it. Every point of the orbit is reached by the same number of elements of $\mathcal{G}$, so the average of D over the orbit's points equals the average of $D(gv)$ over $g \in \mathcal{G}$, which is $R(v)$. □

The argument is the standard one for group-invariant permutation tests (Hemerik and Goeman, 2018, §2, Theorem 1).

This is the general form of the centering the paper states for the randomization benchmark, and of the exact centering of the permutation level for hard units. The paper's reference is the randomization benchmark's estimate of R from its draws. In a world without separation at the real configuration, its mean excesses are 0.0008 under equal weights and 0.0009 under event weights. A redraw of repository-years gives a persistent repository a new vector every year, so it is not a permutation of the objects' vectors, and in the same world it shows 0.0634 and 0.1084. A reference that redraws both channels from the firm's pooled sample shares is not centered either. With one unit in the smaller channel its bias is exactly $-(1-\sum_c \theta_c^2)/N$, with $N = n_P + n_O$, since the pooled sample shares $\hat{\theta}$ of N independent hard draws have $\mathrm{E}\sum_c \hat{\theta}_c^2 = \sum_c \theta_c^2 + (1-\sum_c \theta_c^2)/N$; in general it is negative in every configuration checked, but the sign is not proved.

Figure A1 sets the observed distance beside the randomization benchmark's reference and the technology-position world, firm-year by firm-year, on the comparison sample.

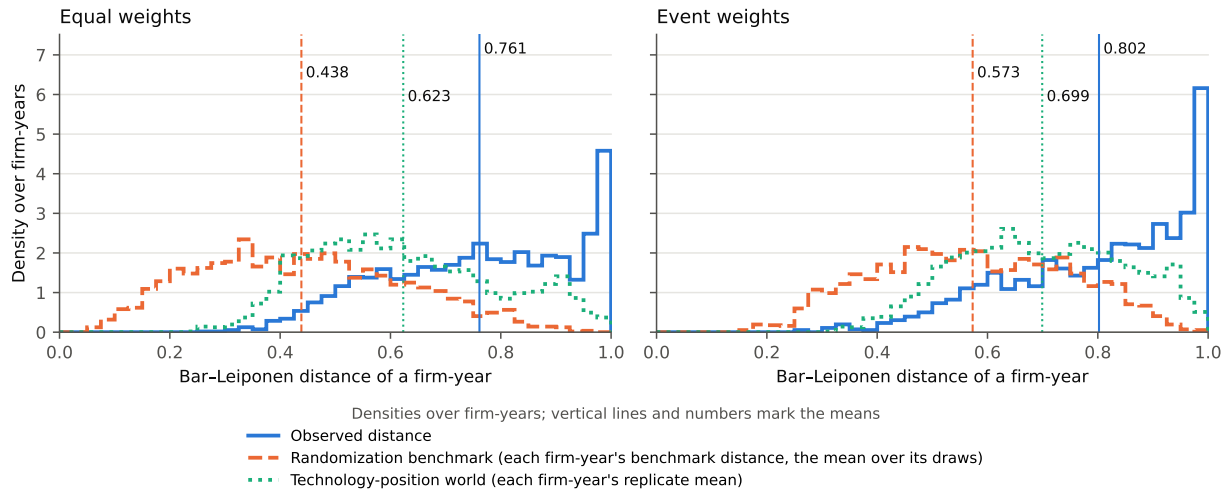


Figure A1: The distance against its reference. Densities over the comparison sample's firm-years of the observed Bar–Leiponen distance (solid), of each firm-year's randomization benchmark reference, its mean over 200 draws (dashed), and of the technology-position world of section 4.4, each firm-year's mean over its replicates (dotted), under equal weights (left) and event weights (right); vertical lines mark the means, which are those of the text and of Table 5.

What the reference measures. By part (i) of the identification proposition (Proposition 2), every statistic of a firm's own units has the same law under a chosen deviation as under technology

norms shifted by it, so a reference built from the firm's own units, the randomization benchmark included, measures technology position and choice together. An excess over it marks a departure from exchangeability of the firm's object vectors across its channels, not its cause. The departure need not be chosen: patents' broader labeling (section A.7) and time trends within a start year are outside the randomization benchmark's law. And the fixed-portfolio world, which fixes each firm's pools, reproduces the excess: 0.3336 and 0.2373, against 0.3304 and 0.2336 observed.

A.9 The Carrington–Troske normalization

Carrington and Troske (1997, §3.1, p. 405) take as the level under random assignment D^* the mean of the index over shuffles that reallocate individuals among units while keeping unit sizes fixed, and define the index of systematic dissimilarity (Carrington and Troske, 1997, §3.2, eq. 6, p. 406):

$$C = \begin{cases} (D - D^*)/(1 - D^*), & D \geq D^*, \\ (D - D^*)/D^*, & D < D^*. \end{cases}$$

In their reading (p. 406), C lies in $[-1, 1]$; $C = 0$ means as separated as random allocation, $C = 1$ complete separation, and a negative value excess evenness; and C tends to D as D^* tends to zero. The range follows directly: above the reference $0 \leq D - D^* \leq 1 - D^*$, and below it $-D^* \leq D - D^* < 0$.

Applied to the Bar–Leiponen distance, $D^* = \mu$ is the mean of the reference's draws, and each distance keeps its own reference: the equal-weight D^A with the equal-weight benchmark, the main D_{it} with the event-weight one, never cross-paired. An endpoint rule was fixed before any reference mean was seen: where $\mu = 1$ the index is undefined, and the row is reported and left out; where $\mu = 0$, $C = D$; and a row whose headroom, $1 - \mu$ above the reference or μ below it, is under 0.05, or under two Monte Carlo standard errors of μ, is flagged, kept, and every summary is repeated without it. On the 2,472 firm-years of the comparison sample every row has an index. Against the randomization benchmark the mean index is 0.6127 under equal weights and 0.5843 under event weights. Against the unit-year reallocation of the firm's own units (section 4.2), which keeps the firm's counts and draws both channels from the same technologies, so that it is an identical-technology reference, but is not centered when repositories persist, it is 0.67 under equal weights and 0.6844 under event weights.

Lemma 11 (The index at an exact null)**.** *Let $\hat{D}$ have mean $\mu \in (0, 1)$ under the reference's own law, and let $\delta = \frac{1}{2}\mathrm{E}|\hat{D} - \mu|$. Then*

$$\mathrm{E}\, C = \frac{\delta\,(2\mu - 1)}{\mu(1 - \mu)},$$

which, unless $\hat{D} = \mu$ with probability one, is positive when $\mu > \frac{1}{2}$ and negative when $\mu < \frac{1}{2}$, and zero at $\mu = \frac{1}{2}$. The one-branch ratio $(\hat{D} - \mu)/(1 - \mu)$ has mean zero.

Proof. Since $\mathrm{E}(\hat{D} - \mu) = 0$, the positive and negative parts of $\hat{D} - \mu$ have equal means, each δ. The upper branch divides the positive part by $1 - \mu$ and the lower branch the negative part by μ, so $\mathrm{E}C = \delta/(1 - \mu) - \delta/\mu = \delta(2\mu - 1)/(\mu(1 - \mu))$. The one-branch ratio is linear in $\hat{D}$. □

In two exact examples, hard units pooled with cell counts $(1, 1, 2, 2)$ and three patents give $\mu = 3/5$ and $\mathrm{E}C = 4/45$, and cell counts $(4, 4)$ with four patents give $\mu = 9/35$ and $\mathrm{E}C = -153/455$. Small portfolios have high levels under random assignment (Lemma 8), so even at an exact null the two-branch index carries a mechanical gradient in the counts. The index is therefore a descriptive normalization, not a test, and a reference's own draws need not average zero on it; where a centered excess is needed, the one-branch ratio is the one to use. Two further limits apply. Carrington and Troske work with whole individuals in fixed units, while the paper's objects are fractional and event-weighted, so the law that generates μ needs its own justification, which equation (6) does not supply; and the index inherits its reference's centering. The paper reports the index, if at all, beside the excess $D - \mu$, as an optional normalization.

A.10 The excess by part

Lemma 12 (The excess over a reference splits exactly)**.** *For each firm-year, let* OUT *and* IN *be the parts of Lemma 3 for the observed configuration, and* OUT_b *and* IN_b *those of reference draw* $b = 1, \ldots, B$*, each computed with that configuration's own patent support. For any weights over firm-years, write* M *for the weighted mean. Then*

$$\mathrm{M}[D] - \mathrm{M}\Big[\tfrac{1}{B}\sum_b D_b\Big] = \Big(\mathrm{M}[\mathrm{OUT}] - \mathrm{M}\Big[\tfrac{1}{B}\sum_b \mathrm{OUT}_b\Big]\Big) + \Big(\mathrm{M}[\mathrm{IN}] - \mathrm{M}\Big[\tfrac{1}{B}\sum_b \mathrm{IN}_b\Big]\Big).$$

Proof. Lemma 3 gives $D = \mathrm{OUT} + \mathrm{IN}$ for every configuration, observed or drawn, and both the average over draws and M are linear. □

The support must be the draw's own: the randomization benchmark moves patent vectors as well as repository vectors, so a draw's patent support differs from the observed one. The lemma is an identity, not a test: the two parts' excesses sum exactly to the excess. What does not add is everything nonlinear: the shares of firm-years with a positive excess, the reference bands and the p-values are computed part by part. A part's excess reads "more separated than the specified reference" only under the excess's finding rule, with the Holm adjustment across the four readings of two parts under two weights, declared before the split is run. The companion reading is whether the fixed-portfolio world, which reproduces the total excess, reproduces each part.

Table A1 gives the split. The two parts sum to the excess: 0.2081 and 0.1223 make 0.3304 under equal weights, and 0.2001 and 0.0335 make 0.2336 under event weights. With each firm-year counted once, most of the mean excess lies outside the support: 63% of it under equal weights and 85.7% under event weights. Weighted by each firm-year's code mass the outside share is 18.12% under equal weights and 6.14% under event weights, so by mass most of the excess lies inside. The four excesses, two parts under two weightings, are findings under the excess's finding rule, with Holm across the four: each part is more separated than the specified reference under both weights. More of a firm's open-source mass lies in main groups where it holds no patents than reallocating its own objects produces, and where the two channels share a main group, the mismatch within it is

Table A1: The excess over the randomization benchmark, by part

Part	Weights	Observed	Benchmark [95% band]	Excess (SE)	Above	Fixed-portfolio world [range]
Outside	Equal	0.3604	0.1523 [0.1477, 0.1564]	0.2081 (0.0098)	81.7%	0.2039 [0.2007, 0.2072]
Within	Equal	0.4041	0.2818 [0.2779, 0.2856]	0.1223 (0.0099)	66.5%	0.1296 [0.127, 0.132]
Total	Equal			0.3304		0.3336
Outside	Event	0.3511	0.1511 [0.1428, 0.1603]	0.2001 (0.0111)	65.9%	0.2037 [0.193, 0.2128]
Within	Event	0.4543	0.4208 [0.4111, 0.4286]	0.0335 (0.0107)	60%	0.0336 [0.0268, 0.0421]
Total	Event			0.2336		0.2373

Means over the comparison sample's 2,472 firm-years. Outside: the open-source mass outside the patent support; within: the mismatch inside it, each configuration with its own support. The randomization benchmark column is the mean of the draws' means with their 2.5th and 97.5th percentiles; the excess is the mean firm-year excess with firm-clustered standard errors; above is the share of firm-years whose part exceeds its own reference mean. The fixed-portfolio world's excess is the mean over its replicates with their 2.5th and 97.5th percentiles.

also larger than the reference's, by much less.

The fixed-portfolio world reproduces the total excesses, 0.3336 and 0.2373. Under event weights it also reproduces each part: the data's excesses lie inside its replicate ranges. Under equal weights it reproduces the outside part. The data's excess there, 0.2081, lies inside the world's replicate range (0.2007 to 0.2072), whose mean is 0.2039. Within the support, the data put a little less of the excess than the world does (0.1223 against 0.1296), just below its narrow replicate range (0.127 to 0.132). Both parts are therefore compatible with fixed portfolios. The world fixes each firm's own pools, so this reading neither tests nor supports the formula of technology norms alone, whose own world is section 4.4's. The split at subgroup depth needs subgroup draws of the benchmark, which do not exist, so it is not run.

A.11 The GitHub classes

Proof of Lemma 16. Let p and o be the firm-year's patent and open-source share vectors, with $o_c = 0$ for $c \notin \text{GH}$, let $\psi = \sum_{c \notin \text{GH}} p_c < 1$, and let $p_c^{\text{GH}} = p_c/(1-\psi)$ for $c \in \text{GH}$, so that $D_{\text{GH}} = D(p^{\text{GH}}, o)$. By Lemma 1, in its form $D = \sum_c (p_c - o_c)^+$,

$$D = \psi + \sum_{c \in \text{GH}} \left((1-\psi)p_c^{\text{GH}} - o_c\right)^+ \;\geq\; \psi,$$

and in its form $D = \sum_c (o_c - p_c)^+$, since $(1-\psi)p_c^{\text{GH}} \leq p_c^{\text{GH}}$,

$$D = \sum_{c \in \text{GH}} \left(o_c - (1-\psi)p_c^{\text{GH}}\right)^+ \;\geq\; \sum_{c \in \text{GH}} \left(o_c - p_c^{\text{GH}}\right)^+ = D_{\text{GH}}.$$

For the upper bound, $(1-\psi)p_c^{\text{GH}} - o_c = (1-\psi)(p_c^{\text{GH}} - o_c) - \psi\, o_c \leq (1-\psi)(p_c^{\text{GH}} - o_c)$, so each term of the first form is at most $(1-\psi)(p_c^{\text{GH}} - o_c)^+$, and $D \leq \psi + (1-\psi)D_{\text{GH}}$. □

Both bounds are attained: the lower one when the open-source portfolio equals the rescaled patent portfolio on GH (then $D_{\text{GH}} = 0$ and $D = \psi$), the upper one when the two portfolios on GH have disjoint supports (then $D_{\text{GH}} = D = 1$). The bound $D \geq \psi$ makes a coefficient on ψ_{it} partly

mechanical when the full distance is the outcome; D_{GH} removes it.

A.12 A distance to another portfolio

The co-location scores of Appendix E.8 measure co-location, not distance. How the distance to a peer group's portfolio moves when a firm moves its code is an identity.

Lemma 13 (A distance to another portfolio)**.** *Let r be a fixed distribution over technologies, such as a peer group's patents or code, and $D(o,r) = \frac{1}{2}\sum_c |o_c - r_c|$.*

(i) If the firm's open-source distribution moves to $o' = (1-\varepsilon)o + \varepsilon r$, with $0 \le \varepsilon \le 1$, then $D(o',r) = (1-\varepsilon)D(o,r)$.

(ii) Moving open-source mass ε from a technology c with $o_c > r_c$ to a technology d with $o_d < r_d$ lowers $D(o,r)$ by exactly ε when $\varepsilon \le \min\{o_c - r_c,\, r_d - o_d\}$; beyond that it lowers it by less than ε, and it raises it once $\varepsilon > (o_c - r_c) + (r_d - o_d)$.

A distance to peers falls when the firm's code moves toward the peers' distribution, or out of technologies where the firm already over-represents the peers into ones where it under-represents them; more code where the peers patent is not enough. With peers' shares $\frac{3}{5}$ and $\frac{2}{5}$ in two technologies and the firm's open-source shares $\frac{7}{10}$ and $\frac{3}{10}$, the distance is $\frac{1}{10}$; moving a tenth of the firm's code into the first technology raises its co-location with the peers and raises the distance to $\frac{1}{5}$, because the firm already held more than its share there. So a positive association with peers' patent presence does not fix the sign of a distance to the peers.

Proof of Lemma 13. (i) $o' - r = (1-\varepsilon)o + \varepsilon r - r = (1-\varepsilon)(o-r)$, so $D(o',r) = \frac{1}{2}\sum_c (1-\varepsilon)|o_c - r_c| = (1-\varepsilon)D(o,r)$.

(ii) Let $a = o_c - r_c > 0$ and $b = r_d - o_d > 0$. The move changes only the terms of c and d, from $a+b$ to $|a-\varepsilon| + |b-\varepsilon|$. If $\varepsilon \le \min\{a,b\}$ the sum falls by 2ε; if $\min\{a,b\} < \varepsilon \le \max\{a,b\}$ it falls by $2\min\{a,b\} < 2\varepsilon$; and if $\varepsilon > \max\{a,b\}$ it changes by $2(\varepsilon - a - b)$, which is above -2ε and positive once $\varepsilon > a + b$. Halving gives the statement. □

The example in the main text: with $r = (\frac{3}{5}, \frac{2}{5})$ and $o = (\frac{7}{10}, \frac{3}{10})$, $D(o,r) = \frac{1}{10}$; moving $\frac{1}{10}$ into the first technology gives $o' = (\frac{4}{5}, \frac{1}{5})$ and $D(o',r) = \frac{1}{5}$, while the covariance of o' with the centered r rises from $\frac{1}{25}$ to $\frac{3}{50}$. A coefficient on an exposure alone therefore fixes no sign for the distance to the peers.

A.13 The four portfolios and the six distances

A firm's industry gives two more portfolios beside the firm's own. The industry is the firm's NAICS 4-digit industry; where the 4-digit industry-year holds fewer than five firms with linked code, it is the 3-digit industry, and the 2-digit industry only where no 3-digit industry is found. Membership is fixed in the year before, the firm itself is left out, and the industry's content is pooled over the firm's own five-year window. Write p_G for the share vector of the industry's patents and o_G for

that of its code, built as p and o are, with the members' masses summed. The four portfolios give six distances: the firm's own, $D(p, o)$; the industry's own, $D(p_G, o_G)$; and four from the firm to its industry, $D(o, o_G)$, $D(o, p_G)$, $D(p, o_G)$ and $D(p, p_G)$. Appendix F puts two other groups in the industry's place: the firms in the top third of the industry-year by R&D rank among US-dollar reporters, and all other owners.

Lemma 14 (The six distances are tied)**.** *(i) Each distance is at most the sum of the other two on any triangle of the four portfolios. In particular,* $|D(o, p_G) - D(p, p_G)| \leq D(p, o)$ *and* $|D(p, o_G) - D(o, o_G)| \leq D(p, o)$.

(ii) $D(p, o)$ *depends only on the firm's portfolios, and* $D(p_G, o_G)$ *only on the industry's. The firm's portfolios enter the other four distances and never the industry's own.*

(iii) If the firm's code moves toward the industry's code, $o' = (1-\varepsilon)o + \varepsilon o_G$ *with* $0 \leq \varepsilon \leq 1$*, then*

$$D(o', o_G) = (1-\varepsilon)D(o, o_G),$$
$$D(o', p_G) \leq (1-\varepsilon)D(o, p_G) + \varepsilon D(p_G, o_G), \qquad D(p, o') \leq (1-\varepsilon)D(p, o) + \varepsilon D(p, o_G).$$

With the channels exchanged, for $p' = (1-\varepsilon)p + \varepsilon p_G$*:* $D(p', p_G) = (1-\varepsilon)D(p, p_G)$*,* $D(p', o_G) \leq (1-\varepsilon)D(p, o_G) + \varepsilon D(p_G, o_G)$ *and* $D(p', o) \leq (1-\varepsilon)D(p, o) + \varepsilon D(o, p_G)$.

(iv) If a group's portfolio on one channel pools two subgroups, $x = wx_A + (1-w)x_B$*, with* w *the first subgroup's share of the channel's mass, then for any portfolio* y *of the firm*

$$|D(y, x) - D(y, x_A)| \leq (1-w)D(x_A, x_B), \qquad D(y, x) \leq wD(y, x_A) + (1-w)D(y, x_B),$$
$$|D(y, x_A) - D(y, x_B)| \leq D(x_A, x_B).$$

(v) If a patent portfolio of every owner holds the firm's patents, $a = sp + (1-s)r$*, with* s *the firm's share of its mass and* r *the portfolio of every other owner, then* $D(p, a) = (1-s)D(p, r)$ *and* $D(o, a) \leq sD(p, o) + (1-s)D(o, r)$*. On the code side the same holds with* p *and* o *exchanged and the firm's share of the code mass in place of* s.

Proof. By Remark 1, D is a metric, which gives (i). Part (ii) holds by construction, since the industry's portfolios leave the firm out. For the rest, $D(x, y) = \frac{1}{2}\|x - y\|_1$ scales with $x - y$ and is convex in each argument. In (iii), $o' - o_G = (1-\varepsilon)(o - o_G)$ gives the identity, as in Lemma 13(i), and convexity in the first argument gives the two bounds; the exchanged case is the same. In (iv), $x - x_A = (1-w)(x_B - x_A)$, so $D(x, x_A) = (1-w)D(x_A, x_B)$, and the triangle inequality gives the first bound; convexity gives the second, and the triangle inequality the third. In (v), $p - a = (1-s)(p - r)$ gives the identity, and convexity gives the bound. □

Each part has a use in reading the distances. By (i), a reading of one distance bounds others, so each is read beside the others. By (ii), nothing on the firm's side moves the industry's own distance. By (iii), a move of the firm's code toward the industry's code lowers the distance between its code

and that code exactly in proportion, but it is sure to lower the firm's own distance only where $D(p, o_G) < D(p, o)$. By (iv), the firm's distances to two subgroups differ by at most the distance between the subgroups' portfolios, and a pooled group's distance lies within $(1 - w)D(x_A, x_B)$ of its heavier part's. By (v), a pool that holds the firm's own patents sits nearer them by exactly the factor $1 - s$, so every reading against all owners is made with the firm left out.

A.14 The references for the four portfolios

Every distance among the four portfolios is read against a reference that holds the counts it depends on. Three references are used. The pools of the industry-pool benchmark and of the within-industry benchmark hold the firm's own objects: over the grouped firm-years the firm holds, at the 90th percentile, 23.5% of its industry-year's patents and 35.9% of its code under event weights, and more than half of its industry-year's code in 7.6% of firm-years (section 5.5).

The randomization benchmark against fixed portfolios. The firm's own objects are reassigned as the randomization benchmark reassigns them, within the firm's start-year strata, while the other portfolios are held as observed. Each draw gives the four distances from the firm to its industry. Against all other owners, the same draw is read against the portfolios of every other owner, with the firm left out: every patent, whoever holds it, and every linked repository, each counted once. A departure reads the firm's channel allocation given its own objects. For the distances to fixed portfolios the randomization benchmark takes 999 draws per firm.

The industry-pool benchmark. For each industry-year, the pool holds the window's objects of the industry's members and of the firm. Each patent and each repository is a slot that carries its owner, its channel and its weight: one for a patent, and for a repository one under equal weights and its window events under event weights. Within each channel and start year, the objects' vectors are reassigned uniformly among the slots; owners, channels and weights stay with the slots. The firm's portfolios and the industry's, with the firm left out, are rebuilt from each draw. The industry-pool benchmark holds each firm's counts by channel, and with them the small-unit bias: the fewer objects a portfolio holds, the further it lies from any pool. It redraws the firm's mix within each channel, so a departure reads the firm's portfolio in a channel against a draw of the same size from that channel's pool. The pool holds the firm's own objects.

The within-industry benchmark. The same pool's objects, patents and repositories together, are reassigned within start-year strata across channels and owners; slots keep their owner, channel and weight, and the four portfolios are read from each draw's slots. Of the three references, only this one crosses channels. A departure of the industry's own distance reads the industry's channel allocation beyond random assignment. The industry-pool benchmark and this benchmark take 499 draws per industry-year.

Centering. Each reference is a uniform draw from a group of permutations of vectors among slots: for the randomization benchmark, the firm's objects within a start year; for the industry-pool benchmark, every owner's slots within a channel and start year; for the within-industry benchmark, every owner's slots within a start year, across channels. The argument of Lemma 10 uses nothing about the statistic but that it is computed from the slots, so it applies to each of the six distances: under a law of the vectors that the reference's permutations leave invariant, the expected excess is zero.

The test. For each firm-year, the excess is the observed distance minus the reference's mean over its draws. A firm-year counts as above the reference when its distance exceeds the mean of its observed value and its draws, as one half at a tie, and as not above otherwise. The share of firm-years above, S_0, is ranked among the same share at each of the B draws, $S_1, \ldots, S_B$, each computed against the same center, and the two-sided p is

$$\min\Big\{1,\ \frac{2\min\big(\#\{k : S_k \geq S_0\},\ \#\{k : S_k \leq S_0\}\big)}{B+1}\Big\}, \qquad k = 0, \ldots, B.$$

The share's mean under a reference is not one half, so the share is read against the draws, not against one half. For the randomization benchmark against fixed portfolios the test pools the firm-years and is Holm-adjusted across the two weightings. The industry-pool benchmark and the within-industry benchmark draw each industry-year on its own, so the test runs within each year, a cross-section with each firm once, and is Holm-adjusted across a row's two weightings and five years (the five years alone where only patents enter, since patents carry no repository weights); the tables give, under a row's weighting, the smallest of the five years' adjusted p, so a value below 5 percent is a rejection in at least one year, and the share above pooled over the years as a description.

Size. The tests reject less often than their nominal level overall in worlds without separation, built on 30 firm-years of the sample drawn at random. For the randomization benchmark, each firm's object vectors are exchangeable across its channels; for the two industry references, every unit of an industry-year draws its technologies from the industry's pooled mix, with the industry's members, counts by channel, weights and timing held. Over 100 such worlds of 199 draws each, the share test rejected at 5 percent in 127 of 3,100 tests of the randomization benchmark against fixed portfolios, 22 of 2,200 of the industry-pool benchmark and 17 of 2,200 of the within-industry benchmark: rates of 4.1%, 1% and 0.8%, with Wilson 95 percent intervals of 3.5% to 4.9%, 0.7% to 1.5% and 0.5% to 1.2%. The seven of the randomization benchmark's statistics that use the industry's leaders among US-dollar reporters as the group reject in 34 of 700 tests, 4.9% (3.5% to 6.7%), below the nominal level. Its other statistics reject in 3.9%. The tests of one replicate share one world and one set of permutations, so the Wilson intervals understate the simulation error. In such worlds the industry references' tests are conservative, and the randomization benchmark's tests reject at about half the nominal level.

A.15 The floors behind the readings against the firm's own objects

When only the firm's own objects are reassigned, the part of a reading that depends on which objects the firm made code is bounded by the observed distances. Let the firm-year hold n_P patents and n_O repositories, with $\pi_P = n_P/(n_P + n_O)$ and $\pi_O = 1 - \pi_P$, and let o^A be its code portfolio under equal weights; the patent portfolio p is the same under both weightings. The randomization benchmark permutes the firm's object vectors within start-year strata, each object keeping its channel and weight. The vector a slot receives then has mean u_s, the mean vector of the firm's objects that start in the slot's year s, so the drawn portfolio of channel $x \in \{p, o\}$ has mean $\bar{u}_x$, its slots' weighted average of the u_s. For a fixed portfolio q, write the excess $\Delta_x(q) = D(x, q) - \mathrm{E}\, D(x^*, q)$ and the sampling term $S_x(q) = \mathrm{E}\, D(x^*, q) - D(\bar{u}_x, q)$.

Lemma 15 (The part that depends on the assignment)**.** *(i) $S_x(q) \geq 0$, and $\Delta_x(q) = \big[D(x, q) - D(\bar{u}_x, q)\big] - S_x(q)$, where the bracket lies within $\pm D(x, \bar{u}_x)$.*

(ii) Without strata, $\bar{u}_p = \bar{u}_o = u = \pi_P p + \pi_O o^A$, the union of the firm's objects, each counted once, under either weighting. Under equal weights $D(p, u) = \pi_O D(p, o)$ and $D(o, u) = \pi_P D(p, o)$, and against the industry's portfolios

$$\begin{aligned} \Delta_o(p_G) + S_o(p_G) &\geq \pi_P\big[D(o, p_G) - D(p, p_G)\big], \\ \Delta_p(o_G) + S_p(o_G) &\geq \pi_O\big[D(p, o_G) - D(o, o_G)\big], \\ \Delta_o(o_G) + S_o(o_G) &\leq D(o, o_G) - D(p, o_G) + \pi_O D(p, o). \end{aligned}$$

(iii) With strata, each bound in (ii) moves by at most $\tau_x = D(u, \bar{u}_x)$.

Proof. (i) D is convex, so Jensen's inequality gives $\mathrm{E}\, D(x^*, q) \geq D(\mathrm{E}x^*, q) = D(\bar{u}_x, q)$, that is, $S_x(q) \geq 0$; the identity is the definition, and the triangle inequality bounds the bracket. (ii) Without strata every slot receives every vector with the same probability, so each drawn portfolio, an average of its slots' vectors with fixed weights summing to one, has mean u. Under equal weights $p - u = \pi_O(p - o)$ and $o - u = \pi_P(o - p)$, which give the two distances. Convexity gives $D(u, p_G) \leq \pi_P D(p, p_G) + \pi_O D(o, p_G)$, so the bracket for $x = o$, $q = p_G$ is at least $(1 - \pi_O)D(o, p_G) - \pi_P D(p, p_G)$; likewise for $x = p$, $q = o_G$. For $x = o$, $q = o_G$, the triangle inequality gives $D(u, o_G) \geq D(p, o_G) - D(p, u) = D(p, o_G) - \pi_O D(p, o)$. (iii) The triangle inequality, $|D(\bar{u}_x, q) - D(u, q)| \leq D(u, \bar{u}_x)$. □

Without strata, then, the excess of the distance between the firm's code and the industry's patents is at least the patent share times the amount by which that distance exceeds the patents' own distance to the industry's patents, net of a sampling term that is never negative; likewise for the firm's patents against the industry's code, with the code share. Without the sampling term the inequality is false. The excess of the distance between the firm's code and the industry's code is at most the amount by which that distance falls short of its patents' distance to the industry's code, plus the code share times the firm's own distance, and the sampling term only lowers it. By (i) and (ii), the part of each reading that depends on which objects the firm made code is at most the

other channel's share times the firm's own distance, so it vanishes without the firm's own separation. This holds for every distance from the firm's portfolios to a fixed portfolio, the patents' distance to the industry's patents included, and twice the bound covers a difference of two such readings. No ordering fixes the side of the patent-to-patent reading.

The paper's randomization benchmark has start-year strata, so (iii) applies. The shift τ_x is at most the total-variation distance between the weights that u and $\bar{u}_x$ put on the firm's objects. When a window holds all of the firm's objects of its start years, it is the other channel's share times the distance between the start-year mixes of the window's patents and code. The records do not compute it, since it needs the firm's objects by start year and window. The bounds are for equal weights. Under event weights the floor of the code against the industry's patents holds in its general form, $D(o, p_G) - \pi_P D(p, p_G) - \pi_O D(o^A, p_G)$, while the code-to-code and patents-to-code bounds need distances the records do not hold. Appendix F.4 describes how the excesses sit beside the bounds.

B Regression tables

This appendix gives the full tables behind the compact tables of the main text. Each table names the table it extends. Every number comes from a recorded result; a cell marked — has no recorded value. Standard errors are clustered by parent firm (Cameron and Miller, 2015) unless a note says otherwise, and every comparison is an association against a named reference. Holm p-values use the sequentially rejective procedure of Holm (1979). The size comparison (Table B41) and the GitHub classes' rows that control for size (Table B12) are computed among US-dollar reporters, the Compustat North America rows reported in US dollars, 1,743 firm-years in the size comparison, and no finding label changes. So are the regressions with a distance or the firm's open-source activity on the left (Appendix B.9 and Table B6): log assets is kept on 2,258 of the 3,064 firm-years with assets, and the leader indicator and the R&D rank are ranked among US-dollar reporters. So are the size rows of Table B3 and the size columns of the GitHub classes' coefficient tables (Tables B10, B11 and B13). For 213 firms without a Compustat North America row the panel reports assets in local currency (799 firm-years), and those firm-years leave every column with log assets.

B.1 The count slope

Table B2 gives the count slope of Appendix 5.2 on four fixed-effect sets, against the randomization benchmark and the fixed-portfolio world; the tables after it extend it.

Table B1 repeats the two main specifications of Table 7.

Table B1: The distance, the benchmark distance and the excess, and the firm's counts, in the two main specifications

	(1)	(2)
Panel A. Observed distance		
Log open-source projects	−0.038***	−0.015***
	(0.004)	(0.005)
Log events per project	0.004	0.006***
	(0.004)	(0.003)
Log patents	0.00039	−0.001
	(0.003)	(0.005)
Patent share outside	0.251***	
open-source classes	(0.021)	
Panel B. Benchmark distance (mean over the benchmark's 200 draws)		
Log open-source projects	−0.071***	−0.066***
	(0.004)	(0.006)
Log events per project	0.019***	0.012***
	(0.004)	(0.004)
Log patents	−0.005	−0.011**
	(0.003)	(0.006)
Patent share outside	0.177***	
open-source classes	(0.027)	
Panel C. Excess over the randomization benchmark (observed minus benchmark distance)		
Log open-source projects	0.034***	0.051***
	(0.004)	(0.007)
Same, without	0.032***	
the patent share	(0.004)	
Log events per project	−0.015***	−0.006
	(0.004)	(0.005)
Log patents	0.005	0.009
	(0.003)	(0.006)
Patent share outside	0.074***	
open-source classes	(0.026)	
Firm fixed effects	No	Yes
Year fixed effects	No	Yes
Industry-by-year fixed effects	Yes	No
Observations	2,447	2,447
Firms	589	589

Notes: The table repeats columns 4 and 6 of Table 7, the two specifications that the text discusses, with the distances under event weights. Column 1 compares firms in the same industry-year, with industry-by-year fixed effects and the patent share outside open-source classes. Column 2 compares a firm with itself over time, with firm and year fixed effects. The outcome is the observed distance in Panel A, the benchmark distance (the mean distance over the benchmark's draws) in Panel B, and the excess (the observed distance minus the benchmark distance) in Panel C. Within each column, the three panels use the same sample and regressors, so each coefficient in Panel C equals the one in Panel A minus the one in Panel B. Each count covers the five-year window that ends the year before the outcome's window ends and enters as the log of one plus the count. Standard errors clustered by firm are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from standard errors clustered by firm.

What the count slope rests on. Within firms the count slope mostly reflects how many units are drawn, and across firms about half of it survives at a common depth. Where the equal-weight distance is computed from the same number of patents and repositories in every firm-year, the slope with firm and year effects is −0.0031 at 5 units, against −0.0204 for the full equal-weight distance on the same firm-years; across firms it is −0.026 against −0.0637. The content of firms with similar patent portfolios reproduces it: when each repository takes the content of repositories from the firm's 20 nearest patent neighbors, by the cosine of Jaffe (1986, eq. 1, p. 987), the slope is

−0.061 against −0.0589 observed (p 0.468 for the difference). It depends on the window: on the 232 firms present under every window it is 0.0608 under event weights on the five-year window and 0.024 on single years, keeping its sign. Within two-digit NAICS codes the slope is flatter, since industry composition takes 19% of it, mostly as a difference in levels under random assignment; and within the classes that carry code, at a given non-GitHub share, the slope lies 0.0344 above the randomization benchmark's (Table B8). The distance is bounded: at main group, 209 of the 3,291 firm-years have a distance of one, and at subgroup depth 8.26% do.

Table B2: The count slope of the distance, against the randomization benchmark

	(1)	(2)	(3)	(4)
Fixed effects	None	Firm	Firm and year	NAICS-2 and year
Count slope β_{repo}, observed	−0.0589	−0.0172	−0.0152	−0.0477
	(0.0041)	(0.0043)	(0.0047)	(0.0044)
The benchmark's slope	−0.0882	−0.0489	−0.0668	−0.0794
Difference, observed minus benchmark	0.0293	0.0318	0.0516	0.0317
	(0.0038)	(0.006)	(0.0072)	(0.0041)
Fixed-portfolio world: difference	0.0295	0.0344	0.0463	0.031
Firm-years	2,472	2,472	2,472	2,409
Parent firms	595	595	595	580

Equation (4). The outcome is the distance under event weights over the five-year window that ends in year t; the regressors are over the five-year window that ends in $t-1$. Every column also includes log events per repository and log patents. Least squares; standard errors clustered by parent firm (Cameron and Miller, 2015) in parentheses. The randomization benchmark's slope is the mean over 200 draws; the difference is the slope of the observed distance minus the randomization benchmark's mean. The fixed-portfolio world's row is its mean difference over 40 replicates. Column 4 keeps the firm-years with a two-digit NAICS code. Columns 1 and 4 compare firms; column 3 is the comparison sample. Table B3 extends this table to equal weights and to the other fixed-effect sets.

Table B4 gives the records behind the reading of Table B2. At a common depth d, the equal-weight distance is computed from d patents and d repositories drawn in every firm-year, so the number of units no longer moves its expected value. The content profiles give each repository the content of other repositories: of the firm's nearest patent neighbors, of firms in its industry, or its own content at other dates.

B.2 Open-source activity, position and the industry's totals

Between firms, firms higher in their industry's R&D rank publish more code, as do leaders at about two standard errors, and firms publish more where their industry's other members hold more code and less where they hold more patents; within a firm over time, the position associations do not show. The regressions put the firm's open-source activity on the left,

$$y_{it} = \beta\, w_{i,t-1} + \boldsymbol{\gamma}^{\top}\mathbf{z}_{i,t-1} + \phi_{it} + \varepsilon_{it}, \tag{3}$$

with $y_{it} = \log(1 + \text{the firm's open-source repositories in the window ending } t)$, $w_{i,t-1}$ the firm's position in its industry-year (the leader indicator, the top third by R&D rank, or the R&D rank's

Table B3: The count slope on every fixed-effect set, under both weights (extends Table B2)

Fixed effects and controls	Observed (SE)	Benchmark	Difference (SE)	Fixed-portfolio world	Firm-years
A. Event weights: the distance D_{it}					
None	−0.0589 (0.0041)	−0.0882	0.0293 (0.0038)	0.0295 [0.0272, 0.0324]	2,472
Year	−0.0591 (0.0041)	−0.0884	0.0294 (0.0038)	—	2,472
Firm	−0.0172 (0.0043)	−0.0489	0.0318 (0.006)	0.0344 [0.0293, 0.0395]	2,472
Firm and year	−0.0152 (0.0047)	−0.0668	0.0516 (0.0072)	0.0463 [0.0407, 0.0517]	2,472
Firm and year; log assets, R&D intensity	−0.0157 (0.006)	−0.0593	0.0436 (0.009)	0.0404 [0.0328, 0.0458]	1,478
Firm and year, on the previous row's sample	−0.0158 (0.006)	−0.0589	0.0431 (0.0091)	—	1,478
NAICS-2 and year	−0.0477 (0.0044)	−0.0794	0.0317 (0.0041)	0.031	2,409
NAICS-3 and year	−0.042 (0.0044)	−0.0742	0.0322 (0.0043)	—	2,409
B. Equal weights: each unit counts once					
None	−0.0637 (0.0044)	−0.0863	0.0226 (0.0039)	0.0255 [0.0245, 0.0265]	2,472
Year	−0.0637 (0.0044)	−0.0865	0.0228 (0.0039)	—	2,472
Firm	−0.0261 (0.0042)	−0.0599	0.0338 (0.0058)	0.0452 [0.0408, 0.0489]	2,472
Firm and year	−0.0204 (0.0048)	−0.0832	0.0628 (0.0069)	0.0598 [0.056, 0.0652]	2,472
Firm and year; log assets, R&D intensity	−0.016 (0.0066)	−0.0773	0.0612 (0.0093)	0.0567 [0.0512, 0.0641]	1,478
Firm and year, on the previous row's sample	−0.0173 (0.0066)	−0.0771	0.0598 (0.0093)	—	1,478
NAICS-2 and year	−0.0501 (0.0045)	−0.0756	0.0255 (0.0041)	0.0276	2,409
NAICS-3 and year	−0.0423 (0.0044)	−0.0679	0.0255 (0.0042)	—	2,409

Equation (4): the coefficient β_{repo} on the lagged log repository count, with lagged log events per repository and log patents in every row. The randomization benchmark's slope is the mean over the benchmark's draws; the difference is the slope of the observed distance minus the randomization benchmark's mean. The fixed-portfolio world gives its mean difference over its replicates, with the replicates' 2.5th to 97.5th percentiles in brackets. The first four rows of each panel use 2,472 firm-years of 595 parent firms; the two rows with size and R&D use the 1,478 firm-years of 360 firms whose assets are reported in US dollars; the NAICS rows use the 2,409 firm-years of 580 firms with a two-digit code (the three-digit code is filled within each firm). The coefficients on the two controls are not reported for these samples; Tables B22 and B23 give every coefficient on the rows of the GitHub-class analysis. Without firm effects a row compares firms with one another.

Table B4: Within and across firms: the count slope at a common depth and under content profiles (extends Table B2)

A. Common depth, equal weights Depth (units per channel)	5	10	20
Within firms (firm and year effects): common depth	−0.0031	−0.0012	−0.0071
full distance, same firm-years	−0.0204	−0.0152	−0.0174
Across firms (no fixed effects): common depth (SE)	−0.026 (0.0033)	−0.0364 (0.0042)	−0.0384 (0.0057)
full distance, same firm-years	−0.0637	−0.0667	−0.0625

B. Content profiles, no fixed effects	Event weights	Equal weights
Observed	−0.0589	−0.0637
Nearest patent neighbors' content	−0.061	−0.0628
p for the difference from observed	0.468	0.768
Same industry's content	−0.0537	−0.0577
p for the difference from observed	0.124	0.057
Own content at other dates	−0.0624	−0.0665
p for the difference from observed	0.177	0.001

Panel B's neighbors are the 20 parents whose whole-period patent mix is closest to the firm's. The neighbor profile shows that technology position suffices for the across-firm slope, not that it is the only account: its donor pool is finite and the neighbors' mixes differ from the firm's.

Table B5: The count slope's difference over the randomization benchmark, with firm and year effects, under other windows (extends Table B2, column 3)

	Own sample		Firms present under every window	
Window	Equal weights	Event weights	Equal weights	Event weights
Five-year (main), recorded regressors	0.0614 (0.0069)	0.0517 (0.0068)	0.0735 (0.0099)	0.0608 (0.0088)
Five-year (main)	0.0614 (0.0068)	0.0516 (0.0068)	0.0735 (0.0099)	0.0608 (0.0088)
Three-year	0.0621 (0.0073)	0.0527 (0.0073)	0.0627 (0.0098)	0.0525 (0.0086)
Single year	0.0314 (0.007)	0.03 (0.0062)	0.0242 (0.0096)	0.024 (0.0104)
Single year, both channels non-empty	0.0278 (0.0048)	0.0231 (0.0043)	0.0188 (0.0095)	0.0171 (0.0098)
Three-year blocks, previous block	−0.0016 (0.0101)	0.0164 (0.0112)	−0.0014 (0.0098)	0.0185 (0.0106)
Five-year blocks, long difference	0.0954 (0.0095)	0.0833 (0.0087)	0.1017 (0.0107)	0.0841 (0.0106)
Five-year blocks, cross-section	0.0099 (0.0058)	0.0188 (0.0056)	0.0079 (0.0073)	0.0193 (0.0074)

Firms present under every window, equal weights

	Observed slope	Benchmark's slope	Difference
Five-year (main)	−0.0145	−0.0881	0.0735
Three-year	−0.0184	−0.0811	0.0627
Single year	−0.0178	−0.042	0.0242
Three-year blocks, previous block	−0.0136	−0.0122	−0.0014
Five-year blocks, long difference	−0.0226	−0.1244	0.1017

The difference is the slope of the observed distance minus the randomization benchmark's mean, with firm and year effects. The distance and the lagged regressors are rebuilt the same way under each window from the previous observation's window. The firms present under every window number 232. The long difference uses same-block counts over the two blocks; the cross-section has no fixed effects.

percentile) or the other members' totals in its industry group (their patent mass, their code mass and the number with code), and $\mathbf{z}_{i,t-1}$ the controls (log patents, log assets, R&D intensity and panel age).

Between firms, under year effects (columns 1 and 4), firm-years higher in their industry's R&D rank hold more repositories, 1.045 (0.495), and leaders hold more at about two standard errors, 0.319 (0.171); within industry-years (column 6) the rank's association is within two standard errors of zero (0.7, 0.583). Within a firm over time neither association shows beyond its standard error (−0.004, 0.048; 0.06, 0.328). Between firms, a larger code mass among the other members of the firm's industry goes with more repositories, 0.296 (0.061), and a larger patent mass with fewer, −0.273 (0.062); within a firm the code mass enters at 0.143 (0.062). Position is a predetermined attribute, not an assignment, and the industry's totals say where firms with more repositories sit, not what the industry does to them: a firm's activity and its industry's are measured together, and nothing here separates their effects on each other.

B.3 What the excess does not rest on, in full

The level and the excess do not rest on the five-year window, on duplicated patents, on code without an open-source license, or on patents in technologies where no code is classified. With single-year windows the mean distance is 0.8056, against 0.8113 on the five-year window, and its excess over the randomization benchmark is 0.2793 under equal weights and 0.2043 under event weights, against 0.3175 and 0.2284, each window on its own firm-years. Dropping the 12,844 patents attached to two or more of the comparison sample's firms, on the bridge the distances read, raises the mean by 0.0008, and the excess on the comparison sample is 0.328 and 0.2325 without them,

Table B6: Open-source activity, the firm's position and its industry's totals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Leader: top third by R&D rank ($t-1$)	0.319	−0.004	0.069						
	(0.171)	(0.048)	(0.19)						
R&D rank's percentile ($t-1$)				1.045	0.06	0.7			
				(0.495)	(0.328)	(0.583)			
Log others' patent mass ($t-1$)							−0.273	−0.138	−0.282
							(0.062)	(0.077)	(0.249)
Log others' code mass ($t-1$)							0.296	0.143	0.007
							(0.061)	(0.062)	(0.249)
Log others with code ($t-1$)							0.128	0.152	0.612
							(0.155)	(0.169)	(0.844)
Log patents ($t-1$)	0.082	−0.009	0.141	0.076	−0.01	0.135	0.212	0.064	0.153
	(0.062)	(0.053)	(0.061)	(0.062)	(0.053)	(0.062)	(0.058)	(0.068)	(0.065)
Log assets ($t-1$)	0.23	0.21	0.363	0.197	0.206	0.317	0.278	0.164	0.354
	(0.067)	(0.065)	(0.075)	(0.073)	(0.075)	(0.08)	(0.063)	(0.089)	(0.077)
R&D intensity ($t-1$)	1.852	0.57	1.773	1.624	0.557	1.472	1.131	0.339	1.391
	(1.06)	(0.221)	(1.116)	(1.058)	(0.236)	(1.086)	(1.025)	(0.121)	(1.153)
Panel age	−0.07	—	−0.042	−0.068	—	−0.042	−0.044	—	−0.038
	(0.025)		(0.028)	(0.025)		(0.028)	(0.027)		(0.032)
Firm		Yes			Yes			Yes	
Year	Yes	Yes		Yes	Yes		Yes	Yes	
Industry × year			Yes			Yes			Yes
Firm-years	1,347	1,347	1,347	1,347	1,347	1,347	974	974	974
Firms	333	333	333	333	333	333	278	278	278

The outcome is $\log(1+$ the firm's open-source repositories in the five-year window ending $t)$; no distance enters. Columns (1)–(3): the leader indicator, the top third by R&D rank in the NAICS industry-year at $t-1$; (4)–(6): the R&D rank's percentile; (7)–(9): the other members' totals in the firm's industry group at $t-1$, the firm left out (patent mass, code mass under event weights, and the number with code), on fiscal years 2021–2024. Columns (1), (4), (7) absorb year effects; (2), (5), (8) firm and year effects (panel age absorbed, a dash); (3), (6), (9) industry-by-year cells. Standard errors clustered by firm, in parentheses. Every column of each ladder is in Appendix B.9. A description, not a test.

against 0.3304 and 0.2336 with them. Keeping on the open-source side only the repositories whose current GitHub license is an open-source license keeps 2,338 of the comparison sample's 2,472 firm-years and 53.3% of its event mass. On that code the distance is 0.792 under equal weights and 0.8274 under event weights, and its excess over the randomization benchmark is 0.28 and 0.2022, against 0.3334 and 0.2354 for all repositories on the same firm-years; each is a finding. The license is today's, not the year's, so this splits the open-source side by its current license rather than by its history. Setting aside the patents in technologies where no code is classified, on the 3,278 firm-years with some patent mass in the classes that carry code, the distance on those classes averages 0.7998, against 0.8106 for the full distance (Appendix B.4). On those classes code is also the more concentrated channel, in 84% of firm-years, and the two portfolios share a main group less often than the randomization benchmark makes them, partly because a patent carries more main groups than a repository (Appendix B.4).

Two further references. The first is the stricter redraw of section 4.4. As at technology norms, it selects which items are open-source projects, holding each start year's number, but at odds set by the rate at which the other firms in the same industry put each item's main groups into open-source code. Its mean distance is 0.688 under event weights, with the draws' 2.5th to 97.5th percentiles from 0.6824 to 0.6934, against 0.6864 at technology norms. It reproduces a further 0.7% of the excess (95 percent interval −0.5% to 1.8%), and 0.4% under equal weights (−0.6% to 1.4%). The second is the fixed-portfolio world. Each firm keeps its patents and its open-source code over the whole period, and only the year of each item is drawn again. Its excess is 0.2373 under event weights and 0.3336 under equal weights, against observed excesses of 0.2336 and 0.3304. The excess therefore does not need a firm's portfolios to change over time.

Figure A2 sets the mean observed distance beside the benchmark's draws.

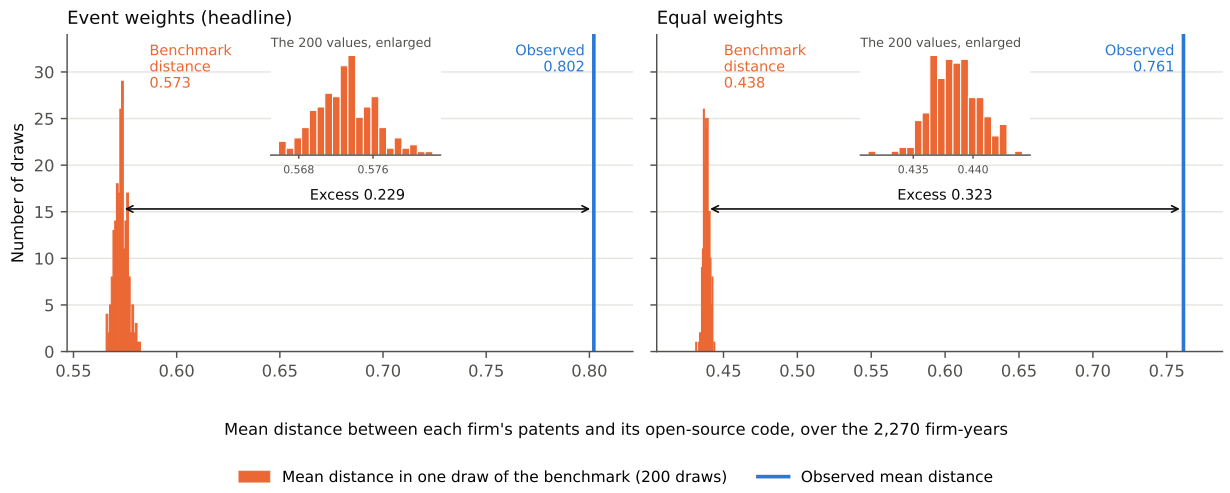


Figure A2: The mean observed distance over firm-years relative to the randomization benchmark

Notes: Each bar reports the number of draws of the benchmark whose mean distance falls in its bin. The inset in each panel enlarges the same values. Each draw reallocates technologies at random among each firm's patents and open-source projects within each start year (the year an item is first seen). Its value is the mean, over the comparison sample's 2,472 firm-years of 595 firms, of the distance D_{it} between each firm's patents and its open-source code over its five-year window. The bars count 200 draws. Their mean is the benchmark distance, the vertical line is the observed mean distance, and the excess is the observed mean distance minus the benchmark distance. Left: event weights, which count each open-source project by its events. The draws lie between 0.564 and 0.579. The benchmark distance is 0.572, the observed mean distance is 0.805, and the excess is 0.234. Right: equal weights, which count each project once. The draws lie between 0.430 and 0.440. The benchmark distance is 0.434, the observed mean distance is 0.764, and the excess is 0.330. The number of draws that reach the observed mean distance is 0 under event weights and 0 under equal weights.

Table B7 splits what technology norms reproduce and the remainder into the parts outside and inside the patent support.

Table B7: Technology norms and the remainder, outside and inside the patent support

	Equal weights	Event weights
Technology norms, outside the patent support	32.8%	34.3%
Technology norms, inside the patent support	82.7%	136.9%
The remainder, outside the patent support	65.2%	63.5%
The remainder, inside the patent support	19.7%	−28.5%

Notes: The rows split the contribution of each step in Table 5 between the excess outside the patent support and the excess inside it, each expressed as a share of the respective part. A share above one, or below zero, means that the step overshoots or undershoots that part.

B.4 Within the GitHub classes

Table B8 gives the count slope on the classes that carry open-source code.

Table B8: The count slope on the classes that carry open-source code, against the randomization benchmark

		Coefficient on log open-source projects			Coefficient on the patent share	
Weights	Fixed effects	Observed	Benchmark [band]	Excess	Benchmark	Excess
Event	None	−0.0416	−0.076 [−0.0807, −0.0705]	0.0344 (0.0042)	0.1798	0.0678 (0.0275)
	Year	−0.0417	−0.0763 [−0.0811, −0.0708]	0.0345 (0.0042)	0.1779	0.0689 (0.0275)
	NAICS-2, year	−0.0399	−0.0745 [−0.0797, −0.0686]	0.0345 (0.0044)	0.2014	0.0193 (0.0314)
	NAICS-3, year	−0.0385	−0.0725 [−0.0777, −0.0666]	0.0341 (0.0045)	0.1773	0.0602 (0.0382)
Equal	None	−0.0422	−0.0706 [−0.0738, −0.0671]	0.0284 (0.0043)	0.2642	0.037 (0.0307)
	Year	−0.0423	−0.0709 [−0.0741, −0.0674]	0.0287 (0.0043)	0.2619	0.0389 (0.0307)
	NAICS-2, year	−0.0408	−0.0689 [−0.0724, −0.0653]	0.0281 (0.0045)	0.2959	−0.037 (0.0359)
	NAICS-3, year	−0.0385	−0.0661 [−0.0694, −0.0624]	0.0275 (0.0045)	0.2462	0.0079 (0.0399)

Notes: The open-source classes are the main groups that carry some firm's open-source code. Each row reports estimates of equation (1) using the distance between the firm's open-source code and its patents in the open-source classes as the outcome. The patents in those classes are rescaled to sum to one. The regressors are the three counts and the patent share outside those classes, from the outcome's own window, and no row has firm effects. NAICS-2 is the two-digit NAICS sector and NAICS-3 the three-digit industry. The sample consists of firm-years with a two-digit NAICS code and some patents in the open-source classes. In every draw, the randomization benchmark keeps both channels within the open-source classes. The benchmark coefficient is the mean over the benchmark's draws. The band in brackets runs from the 2.5th to the 97.5th percentile of the benchmark's draws. The excess column gives the coefficient for the observed outcome minus the benchmark outcome, with its standard error, clustered by firm (Cameron and Miller, 2015), in parentheses. Appendix Table B12 adds the other sets of fixed effects.

The GitHub classes. Some patents sit in technologies where no open source is classified, and four constructions set them aside. The *GitHub classes* GH are the CPC main groups that carry open-source mass in any firm-year window: 723 main groups in 132 subclasses, against 6,074 main groups that the firms' patents touch. They are the classes where the labeler placed linked repositories, not all of GitHub. The *non-GitHub share* ψ_{it} is the share of the firm-year's patent mass outside GH; it has this one meaning throughout. Its mean is 0.1758 and its median 0.0427. It describes classification coverage. The *patent portfolio on the GitHub classes* is the firm-year's patent vector restricted to GH and rescaled to sum to one. D_{GH} is the Bar–Leiponen distance between that portfolio and the open-source portfolio, all of which lies in GH.

Lemma 16 (The distance and the GitHub classes)**.** *If all of a firm-year's open-source mass lies in* GH *and* $\psi < 1$ *is the share of its patent mass outside* GH*, then*

$$\max\{\psi,\, D_{\mathrm{GH}}\} \;\le\; D \;\le\; \psi + (1-\psi)\, D_{\mathrm{GH}}.$$

The proof is in Appendix A.11. The full distance is at least the non-GitHub share, so with the full distance as an outcome a coefficient on ψ_{it} is partly mechanical; and it exceeds D_{GH} by at most $\psi(1 - D_{\mathrm{GH}})$, so a firm-year whose patents sit mostly in the GitHub classes keeps nearly its whole

distance when the rest is set aside. On the 3,278 firm-years with $\psi_{it} < 1$, D_{GH} averages 0.7998 against 0.8106 for the full distance, so setting aside the patents outside the classes that carry code moves the mean distance little. The concentration below and the count slope of Appendix 5.2 use these classes.

Open source is the more concentrated channel. On the GitHub classes, open source is more concentrated than the firm's patents in 84% of firm-years (Table B9). Most of the raw gap is what the randomization benchmark produces: 0.11 of 0.1262. Over pairs of distinct units, an index that does not rise when fewer units are drawn, the gap is 0.0519 against 0.0026 under the randomization benchmark. Overlap is below its level under random assignment: 95.3% of firm-years share a main group, against 99.8% under the randomization benchmark. Part of the difference is classification, since a patent carries more main groups than a repository. Without duplicated patents open source is still the more concentrated channel, in 83.8% of firm-years. No reason for opening signs concentration within the classes where code is placed, so the comparison stays descriptive.

Table B9: Concentration and overlap on the GitHub classes, against the randomization benchmark

A. Concentration on GH *(HHI), event weights*

	Raw		Distinct-unit	
	Observed	Benchmark	Observed	Benchmark
Patents on GH	0.137	0.1148	0.1178	0.0977
Open source on GH	0.2632	0.2248	0.1708	0.1002
Gap, open source minus patents	0.1262	0.11	0.0519	0.0026
Excess, observed minus benchmark	0.0162 (0.0056)		0.0495 (0.0053)	

B. Overlap

	Observed	Benchmark	Difference
Shares a main group	0.9526	0.9985	−0.0459 (0.0077)
Shares a subclass	0.9777	0.9996	−0.0219 (0.0055)
Co-located patent share on GH	0.458	0.7609	−0.3029 (0.0088)

Means over 2,466 firm-years of 594 parent firms; standard errors clustered by parent firm (Cameron and Miller, 2015) in parentheses. The randomization benchmark exchanges technology vectors between the channels within each firm and start year. The distinct-unit HHI drops each unit's overlap with itself, so it does not rise when fewer units are drawn. The co-located patent share is the share of the patent mass on GH that lies in main groups where the firm also has open-source mass. Table B14 gives the full version, and Table B13 the overlap measures' regressions.

The regressions. This subsection gives the regressions within the GitHub classes without firm effects (Tables B8, B22 and B23); Tables B10, B11 and B12 give every column, with every coefficient: the repository count, events per repository, patents, the non-GitHub share ψ_{it}, each fixed-effect set, and the number of observations. The regressions with firm and year effects are for comparison only. With the full distance as the outcome, the coefficient on ψ_{it} is partly mechanical, because the full distance is at least ψ_{it}.

The overlap measures do not move with the repository count the way the distance does

Table B10: The distance within the GitHub classes, D_{GH}, on the count regressors and ψ_{it}: every coefficient (extends Table B22)

Fixed effects	None	Year	NAICS-2, year	NAICS-3, year	NAICS-2, year; size, R&D	NAICS-2, year; same sample	Firm and year
A. D_{GH}, event weights							
$\log(1+n^{\text{repo}})$	−0.0416	−0.0417	−0.0399	−0.0385	−0.0487	−0.0456	−0.0159
	(0.0042)	(0.0042)	(0.0044)	(0.0044)	(0.0053)	(0.0053)	(0.0052)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	0.0008	0.0006	0.0009	0.0046	−0.0023	−0.0015	0.0068
	(0.004)	(0.004)	(0.0041)	(0.0042)	(0.0054)	(0.0052)	(0.0032)
$\log(1+n^{\text{pat}})$	0.0027	0.0027	0.0012	−0.0021	−0.0047	0.0024	−0.0013
	(0.0027)	(0.0027)	(0.0029)	(0.0029)	(0.005)	(0.0039)	(0.0049)
ψ_{it}	0.2476	0.2467	0.2206	0.2374	0.1771	0.2131	0.117
	(0.0174)	(0.0175)	(0.0229)	(0.0257)	(0.0342)	(0.0319)	(0.05)
log assets					0.0107		
					(0.0058)		
R&D intensity					−0.0929		
					(0.0644)		
Constant	0.8911	0.8924	0.8957	0.8844	0.8936	0.918	0.8037
	(0.023)	(0.0231)	(0.0237)	(0.0243)	(0.0395)	(0.0309)	(0.0325)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.352	0.353	0.378	0.419	0.464	0.451	0.956
Within R^2	0.352	0.353	0.236	0.204	0.305	0.288	0.024
B. D_{GH}, equal weights							
$\log(1+n^{\text{repo}})$	−0.0422	−0.0423	−0.0408	−0.0385	−0.0466	−0.0443	−0.0236
	(0.0042)	(0.0042)	(0.0044)	(0.0044)	(0.0053)	(0.0053)	(0.0052)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0047	−0.0047	−0.0035	−0.0004	−0.002	−0.0025	−0.0027
	(0.004)	(0.004)	(0.004)	(0.0041)	(0.005)	(0.005)	(0.0025)
$\log(1+n^{\text{pat}})$	−0.0033	−0.0033	−0.0055	−0.0094	−0.0105	−0.0063	−0.005
	(0.0026)	(0.0026)	(0.0029)	(0.0028)	(0.0048)	(0.0037)	(0.0053)
ψ_{it}	0.3012	0.3008	0.2589	0.2541	0.2424	0.2713	0.1672
	(0.0178)	(0.0179)	(0.0219)	(0.0245)	(0.0343)	(0.0321)	(0.0516)
log assets					0.0053		
					(0.0056)		
R&D intensity					−0.1701		
					(0.0524)		
Constant	0.8962	0.8967	0.9025	0.8971	0.9138	0.9101	0.8458
	(0.0229)	(0.023)	(0.0235)	(0.0234)	(0.0379)	(0.0302)	(0.0355)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.439	0.439	0.472	0.514	0.545	0.531	0.969
Within R^2	0.439	0.439	0.312	0.271	0.385	0.366	0.061

n^{repo}, n^{ev} and n^{pat} are repositories, open-source events and patents over the five-year window that ends in $t-1$; ψ_{it} is from the outcome's own window. The rows are the firm-years with a two-digit NAICS code and $\psi_{it} < 1$; the columns with size and R&D, and the column after them, use the firm-years with those controls. Standard errors clustered by parent firm in parentheses.

Table B11: The full distance on the count regressors and ψ_{it}, on the same rows (extends Table B23)

Fixed effects	None	Year	NAICS-2, year	NAICS-3, year	NAICS-2, year; size, R&D	NAICS-2, year; same sample	Firm and year
A. D_{it}, event weights							
$\log(1+n^{\text{repo}})$	−0.0413	−0.0415	−0.0393	−0.0379	−0.0474	−0.0445	−0.0138
	(0.0041)	(0.0041)	(0.0043)	(0.0043)	(0.0053)	(0.0052)	(0.0048)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0008	−0.0009	−0.0001	0.0036	−0.0031	−0.0023	0.0062
	(0.0039)	(0.0039)	(0.004)	(0.0041)	(0.0053)	(0.0051)	(0.003)
$\log(1+n^{\text{pat}})$	0.0041	0.004	0.002	−0.0009	−0.0035	0.0033	−0.002
	(0.0026)	(0.0026)	(0.0027)	(0.0028)	(0.0049)	(0.0038)	(0.0047)
ψ_{it}	0.2819	0.281	0.2516	0.2697	0.2191	0.2524	0.1782
	(0.0171)	(0.0172)	(0.0219)	(0.0245)	(0.0334)	(0.0313)	(0.0409)
log assets					0.0101		
					(0.0057)		
R&D intensity					−0.0792		
					(0.0633)		
Constant	0.8972	0.8985	0.9007	0.8876	0.8934	0.9178	0.8031
	(0.022)	(0.0221)	(0.0226)	(0.0235)	(0.0388)	(0.0301)	(0.0311)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.404	0.405	0.428	0.463	0.505	0.495	0.961
Within R^2	0.404	0.404	0.266	0.227	0.326	0.312	0.033
B. Equal weights							
$\log(1+n^{\text{repo}})$	−0.0418	−0.0418	−0.0399	−0.0378	−0.0455	−0.0433	−0.019
	(0.0041)	(0.0041)	(0.0043)	(0.0043)	(0.0052)	(0.0052)	(0.0048)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0063	−0.0064	−0.0046	−0.0015	−0.0028	−0.0032	−0.0029
	(0.0039)	(0.0039)	(0.0039)	(0.0039)	(0.005)	(0.0049)	(0.0023)
$\log(1+n^{\text{pat}})$	−0.001	−0.001	−0.0037	−0.0074	−0.0088	−0.0046	−0.0056
	(0.0025)	(0.0025)	(0.0027)	(0.0027)	(0.0046)	(0.0036)	(0.0052)
ψ_{it}	0.347	0.3466	0.3047	0.3052	0.2947	0.3228	0.2487
	(0.0177)	(0.0178)	(0.0215)	(0.0236)	(0.0345)	(0.0321)	(0.0514)
log assets					0.0055		
					(0.0056)		
R&D intensity					−0.1558		
					(0.0526)		
Constant	0.8985	0.899	0.9026	0.8959	0.9075	0.9062	0.8326
	(0.0219)	(0.022)	(0.0225)	(0.0227)	(0.0373)	(0.0293)	(0.034)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.5	0.5	0.529	0.567	0.592	0.58	0.973
Within R^2	0.5	0.5	0.353	0.302	0.412	0.396	0.081

n^{repo}, n^{ev} and n^{pat} are repositories, open-source events and patents over the five-year window that ends in $t-1$; ψ_{it} is from the outcome's own window. The rows are the firm-years with a two-digit NAICS code and $\psi_{it} < 1$; the columns with size and R&D, and the column after them, use the firm-years with those controls. Standard errors clustered by parent firm in parentheses.

Table B12: D_{GH} against the randomization benchmark on every fixed-effect set (extends Table B8)

		Count slope			Coefficient on ψ_{it}	
Weights	Fixed effects	Observed	Benchmark [band]	Difference (SE)	Benchmark	Difference (SE)
Event	None	−0.0416	−0.076 [−0.0807, −0.0705]	0.0344 (0.0042)	0.1798	0.0678 (0.0275)
	Year	−0.0417	−0.0763 [−0.0811, −0.0708]	0.0345 (0.0042)	0.1779	0.0689 (0.0275)
	NAICS-2, year	−0.0399	−0.0745 [−0.0797, −0.0686]	0.0345 (0.0044)	0.2014	0.0193 (0.0314)
	NAICS-3, year	−0.0385	−0.0725 [−0.0777, −0.0666]	0.0341 (0.0045)	0.1773	0.0602 (0.0382)
	NAICS-2, year; size, R&D	−0.0487	−0.0762 [−0.0824, −0.0689]	0.0275 (0.006)	0.2099	−0.0328 (0.0504)
	NAICS-2, year; same sample	−0.0456	−0.0722 [−0.0784, −0.0648]	0.0266 (0.0056)	0.2565	−0.0434 (0.0481)
	Firm and year	−0.0159	−0.0682 [−0.0788, −0.058]	0.0523 (0.0077)	−0.0339	0.1509 (0.0697)
Equal	None	−0.0422	−0.0706 [−0.0738, −0.0671]	0.0284 (0.0043)	0.2642	0.037 (0.0307)
	Year	−0.0423	−0.0709 [−0.0741, −0.0674]	0.0287 (0.0043)	0.2619	0.0389 (0.0307)
	NAICS-2, year	−0.0408	−0.0689 [−0.0724, −0.0653]	0.0281 (0.0045)	0.2959	−0.037 (0.0359)
	NAICS-3, year	−0.0385	−0.0661 [−0.0694, −0.0624]	0.0275 (0.0045)	0.2462	0.0079 (0.0399)
	NAICS-2, year; size, R&D	−0.0466	−0.0688 [−0.0732, −0.0639]	0.0223 (0.0058)	0.3547	−0.1123 (0.0577)
	NAICS-2, year; same sample	−0.0443	−0.0654 [−0.0697, −0.0606]	0.021 (0.0054)	0.396	−0.1247 (0.0539)
	Firm and year	−0.0236	−0.0854 [−0.0938, −0.0766]	0.0618 (0.0075)	0.0702	0.097 (0.0712)

The randomization benchmark restricts both channels to GH in every draw; its slope is the mean over the draws, with the draws' 2.5th to 97.5th percentiles in brackets, and the difference is the slope of the observed minus the outcome under the benchmark. Much of the coefficient on ψ_{it} is what the randomization benchmark produces: the level under random assignment of D_{GH} rises with ψ_{it}, because a firm-year with little patent mass on GH has a thin restricted portfolio.

(Table B13). The two overlap dummies do not move with it; they rise with the patent count, by 0.0234 (SE 0.0047) for the main-group dummy with two-digit NAICS and year effects, and they fall with ψ_{it}. The co-located patent share rises steeply with the repository count, by 0.1156 with the same effects, and mostly as the randomization benchmark's does (0.0976). Its excess over the randomization benchmark, 0.018 (SE 0.0069), runs the other way from the count slope's excess in Table B2: on this measure, firm-years with more repositories place somewhat more of their patent mass on GH beside their own open source than the randomization benchmark does. The two measures differ, since the co-located share weights the patent side and counts any shared main group alike, so the paper reads the excess of the count slope against the distance only, and records this one here.

B.5 The presence association

At the firm–technology–year level, lagged patent presence in a CPC main group goes with a 0.235-percentage-point higher probability of open-source activity there, against an unconditional mean of 0.5062%: small in absolute terms, large against a sparse base rate. The pattern holds in the first year a firm becomes active in a technology, not only in years it continues there. The association is computed on the grid of Appendix C.5, with firm–technology, firm–year and technology–year effects. It is about presence, not about how much: the open-source share of a technology's activity shows no association with earlier patents distinguishable from zero. Appendix E.8 reads the order of the two channels within a technology.

Table B13: The overlap measures on the count regressors and ψ_{it}: every coefficient (extends Table B9, panel B)

Fixed effects	None	Year	NAICS-2, year	NAICS-3, year	NAICS-2, year; size, R&D	NAICS-2, year; same sample	Firm and year
A. Shares a main group (linear probability)							
$\log(1+n^{\text{repo}})$	−0.0078	−0.0079	−0.0046	−0.0017	0.0014	0.002	0.0027
	(0.0049)	(0.0049)	(0.0055)	(0.0054)	(0.0058)	(0.0056)	(0.0085)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	0.0001	0.0001	0.004	0.0009	0.0046	0.0049	0.0203
	(0.0049)	(0.0049)	(0.0047)	(0.0047)	(0.0051)	(0.0049)	(0.0099)
$\log(1+n^{\text{pat}})$	0.0274	0.0274	0.0234	0.0206	0.015	0.0165	0.0067
	(0.0043)	(0.0043)	(0.0047)	(0.0046)	(0.0075)	(0.0058)	(0.0089)
ψ_{it}	−0.4079	−0.4081	−0.4115	−0.4248	−0.2427	−0.2368	−0.1398
	(0.0636)	(0.0637)	(0.0709)	(0.0782)	(0.0935)	(0.0882)	(0.1546)
log assets					0.0023		
					(0.0094)		
R&D intensity					0.0007		
					(0.0928)		
Constant	0.9219	0.9224	0.9095	0.9296	0.8806	0.8886	0.8328
	(0.032)	(0.0322)	(0.033)	(0.0335)	(0.0603)	(0.0417)	(0.078)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.23	0.23	0.253	0.315	0.189	0.189	0.809
Within R^2	0.23	0.23	0.172	0.138	0.094	0.093	0.01
B. Shares a subclass (linear probability)							
$\log(1+n^{\text{repo}})$	−0.0033	−0.0033	−0.0003	0.0007	0.0023	0.0039	−0.0116
	(0.0032)	(0.0032)	(0.0036)	(0.0038)	(0.0045)	(0.0039)	(0.0095)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0068	−0.0068	−0.0036	−0.005	−0.0037	−0.0029	−0.0153
	(0.0045)	(0.0045)	(0.0039)	(0.0041)	(0.0035)	(0.0035)	(0.0123)
$\log(1+n^{\text{pat}})$	0.0138	0.0139	0.0103	0.0108	0.0015	0.0056	0.008
	(0.0035)	(0.0035)	(0.0033)	(0.0036)	(0.0048)	(0.0022)	(0.0046)
ψ_{it}	−0.2366	−0.2367	−0.2146	−0.2432	−0.1287	−0.1116	−0.1313
	(0.0596)	(0.0597)	(0.0531)	(0.0568)	(0.0581)	(0.0507)	(0.0978)
log assets					0.0066		
					(0.0089)		
R&D intensity					0.0013		
					(0.0966)		
Constant	1.0007	1.0007	0.9866	0.9921	0.9515	0.9742	1.0855
	(0.0255)	(0.0257)	(0.0256)	(0.0269)	(0.0536)	(0.0301)	(0.0873)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.16	0.16	0.194	0.286	0.142	0.138	0.847
Within R^2	0.16	0.16	0.097	0.096	0.047	0.042	0.018
C. Co-located patent share on GH							
$\log(1+n^{\text{repo}})$	0.1209	0.1211	0.1156	0.1133	0.1242	0.1212	0.0659
	(0.0059)	(0.0059)	(0.0065)	(0.0065)	(0.0079)	(0.0079)	(0.008)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	0.0193	0.0195	0.0159	0.0109	0.0169	0.0162	0.0044
	(0.0062)	(0.0062)	(0.0063)	(0.0062)	(0.0082)	(0.0081)	(0.0042)
$\log(1+n^{\text{pat}})$	−0.021	−0.021	−0.0162	−0.012	−0.0133	−0.0203	0.0027
	(0.0039)	(0.0039)	(0.0046)	(0.0045)	(0.0083)	(0.0059)	(0.0084)
ψ_{it}	−0.4659	−0.4647	−0.4182	−0.4302	−0.3732	−0.4084	−0.2882
	(0.0297)	(0.0297)	(0.039)	(0.0434)	(0.0591)	(0.0558)	(0.0743)
log assets					−0.0104		
					(0.0102)		
R&D intensity					0.094		
					(0.0942)		
Constant	0.0824	0.0806	0.0892	0.1058	0.121	0.0979	0.226
	(0.0339)	(0.034)	(0.035)	(0.0355)	(0.0675)	(0.0477)	(0.0501)
Firm-years	2,403	2,403	2,403	2,403	1,450	1,450	2,403
Parent firms	579	579	579	579	354	354	579
R^2	0.609	0.61	0.627	0.663	0.678	0.675	0.971
Within R^2	0.609	0.61	0.482	0.451	0.53	0.525	0.101

n^{repo}, n^{ev} and n^{pat} are repositories, open-source events and patents over the five-year window that ends in $t-1$; ψ_{it} is from the outcome's own window. The rows are the firm-years with a two-digit NAICS code and $\psi_{it} < 1$; the columns with size and R&D, and the column after them, use the firm-years with those controls. Least squares; panel A is a linear probability model, read as an average association (Papke and Wooldridge, 2008). Standard errors clustered by parent firm in parentheses.

Table B14: Concentration and overlap on the GitHub classes in full, and the construction's sensitivities (extends Table B9)

A. Concentration (HHI)

	Raw, observed	Raw, benchmark	Distinct-unit, observed	Distinct-unit, benchmark
Patents, full vector	0.1108	0.0983	0.1	0.0842
Patents on GH	0.137	0.1148	0.1178	0.0977
Open source on GH, event weights	0.2632	0.2248	0.1708	0.1002
Open source on GH, equal weights	0.1697	0.1318	0.1488	0.0998
Gap, open source minus patents on GH, event	0.1262	0.11	0.0519	0.0026
Gap, open source minus patents on GH, equal	0.0327	0.017	0.0306	0.0021
Excess, observed minus benchmark (SE), event	0.0162 (0.0056)		0.0495 (0.0053)	
Excess, observed minus benchmark (SE), equal	0.0157 (0.0047)		0.0286 (0.0044)	

B. Overlap

	Observed	Benchmark	Difference (SE)
Shares a main group	0.9526	0.9985	−0.0459 (0.0077)
Shares a subclass	0.9777	0.9996	−0.0219 (0.0055)
Co-located patent share on GH	0.458	0.7609	−0.3029 (0.0088)
Co-located open-source share, equal weights	0.6412	0.8483	−0.2071 (0.0098)
Co-located open-source share, event weights	0.6504	0.8494	−0.199 (0.011)

C. Sensitivities: count slope of D_{GH} and its difference over the randomization benchmark

	Event weights	Equal weights	Difference, event	Difference, equal
Primary, no fixed effects	−0.0416	−0.0422	0.0344	0.0284
Primary, NAICS-2 and year	−0.0399	−0.0408	0.0345	0.0281
Lagged ψ_{it}, no fixed effects	−0.0418	−0.0424	0.0341	0.0282
Lagged ψ_{it}, NAICS-2 and year	−0.0401	−0.041	0.0343	0.028
Leave-one-firm-out GH, no fixed effects	−0.0417	−0.0423	0.0344	0.0285
Leave-one-firm-out GH, NAICS-2 and year	−0.0401	−0.0409	0.0346	0.0282
Per-window GH, no fixed effects	−0.0406	−0.041	0.0354	0.0294
Per-window GH, NAICS-2 and year	−0.0391	−0.0396	0.0356	0.0291

Panels A and B: means over 2,466 firm-years of 594 parent firms. The randomization benchmark exchanges vectors between the channels, so its concentration gap reflects counts and weights, not each channel's own classification; a patent carries more main groups than a repository, so concentration mixes how each channel concentrates with how it is classified. The leave-one-firm-out set does not depend on the firm's own allocation, so its reference is exact.

B.6 A property of the technology or of the firm

Firms differ from one another in how much of their open-source code lies outside their patented technologies, more than the randomization benchmark produces (Table B16, Panel A). The variance of firms' mean outside rates is 0.0937, against 0.0621 under the benchmark, and it lies above its band on every variant of the data in Table B15. In the common-rate world of Panel B, which has no firm term, the dispersion across firms is 0.0248 to 0.0273. The split of the outside indicator's variation between technology-year and firm-year effects, the margin, depends on how patents are dated. With patents dated by grant, technology-year effects explain less of the indicator's variation, relative to firm-year effects, than the randomization benchmark produces. With patents dated by filing, the margin lies just inside its band on all windows and inside it on the windows that end in 2020–2022.

Table B15: The margin and the dispersion across firms on each variant of the data

	The margin			Dispersion across firms		
Variant	Observed	Band	Holm p	Observed	Band	Holm p
Grant date, windows ending 2020	−0.1691	[−0.1628, −0.1312]	0.0199	0.0950	[0.0610, 0.0685]	0.0199
Grant date, all windows	−0.1486	[−0.1475, −0.1265]	0.0498	0.0937	[0.0598, 0.0647]	0.0199
Grant date, without subsidiaries	−0.1548	[−0.1419, −0.1234]	0.0199	0.0945	[0.0592, 0.0640]	0.0199
Filing date, all windows	−0.1131	[−0.1134, −0.0912]	0.0697	0.0854	[0.0542, 0.0591]	0.0199
Filing date, windows ending 2020–2022	−0.1227	[−0.1300, −0.1070]	0.5274	0.0866	[0.0533, 0.0584]	0.0199

The margin is the difference between the shares of the outside indicator's variation that technology-year and firm-year effects explain, and the dispersion across firms is the variance of firms' mean outside rates after removing year means, as in Table B16. Each row is read against the randomization benchmark on its own sample. The band holds the central 95 percent of the benchmark's draws. Holm p is by the plus-one rule (Phipson and Smyth, 2010), across the two statistics. Grant date, windows ending 2020: the 420 firm-years whose five-year windows end in 2020. Without subsidiaries: 2,385 firm-years of 579 firms, with each firm's subsidiaries left out. Filing date, windows ending 2020–2022: the 1,460 firm-years of 542 firms whose windows end in 2020, 2021 or 2022, with patents dated by filing year.

Patents dated by filing. Grant date is the paper's date for patents in the distances, their regressions and the split above. Dating them by filing year changes three of these results. The tests of order and timing in Appendix E.8 date patents by filing year, with grant year as a check (Table B19). The margin is the first of the three results, as Table B15 shows. The second is the within-firm coefficient on log patents in Panel C of Table 7 (column 5). By filing date, the excess rises with the firm's patent count. The coefficient is 0.026 (standard error 0.007) on all windows and 0.028 (standard error 0.007) on the firm-years whose windows end in 2020–2022, so it holds without the last two windows, where recent filings are still censored. By grant date, the coefficient is 0.00018 (standard error 0.006). The third is the check of section 5.2.3 within a firm, against simulated firms that keep one technology mix at every size. On the firm-years whose windows end in 2020–2022, the data's excess rises by 0.0266 per doubling of the open-source projects, against 0.0198 in the simulated firms, whose band runs from 0.0144 to 0.0249. The simulated firms' rise thus falls below the data's on those windows, and the fixed-mix world does not reproduce the slope there (Monte Carlo $p = 0.030$ before Holm's adjustment and 0.060 after it). On all windows by filing date, the

data's rise is 0.0252, inside the simulated firms' band from 0.0195 to 0.0272 ($p = 0.547$).

Table B16 gives the split of Appendix B.6: the outside indicator against the object-level reference, and the common-rate world.

Table B16: Code outside the firm's patent technologies against the object-level reference: dispersion across firms and across technologies

Statistic	Observed	Reference mean	Reference band	Holm p
A. The split of the outside indicator				
Dispersion across firms	0.0937	0.0621	[0.0598, 0.0647]	$\leq$ 0.0 [0.0199]
Dispersion across technologies	0.1261	0.1135	[0.109, 0.1182]	—
B. The common-rate world: one outside rate per technology, no firm term				
Dispersion across firms	0.0248–0.0273	—	high 0.019–0.0196	—
Dispersion across technologies	0.1354	—	midpoint 0.1535	—

The reference is the randomization benchmark (Table 4), 200 draws; the band is the draws' 2.5th to 97.5th percentiles. Holm p is by the rule declared before the test, with its plus-one companion in brackets (Phipson and Smyth, 2010). Panel A: the outside indicator on the technologies where the firm has open-source activity, 52,342 active technology-years in 528 technologies, 2,450 firm-years of 595 firms. The dispersion across firms is the variance of firms' mean outside rates after removing year means. Panel B: the range over 20 worlds, each against its own reference band; the dispersion across firms is above its band in 20 of them. Table B17 runs the same split on the firm's patents.

B.7 The tests of Appendix E.8

The placement relative to nearest patent neighbors and complementors, and the split of the excess, have no tables here: their results are reported in the text and in Appendix A. The order-and-timing test of Appendix E.8 is in Tables B18 and B19.

The co-location scores' peer groups. The three peer groups of Appendix E.8 are built from patents and rebuilt on every draw of the randomization benchmark, so they are never held fixed while the outcomes are redrawn. They are the ten nearest patent neighbors, the firms whose patent technologies are closest to the firm's own; the fifty firms with the most similar patent technologies, weighted by their growth in patent grants and by whether the firm trails them in the technology; and complementors, the product-market peers of Hoberg and Phillips (2016) in the firm's two-digit SIC industry whose patent technologies overlap little with its own, a fraction of the firm-years with the same peer firms shared heavily across groups. Dropping the 12,838 patents attached to two or more of the comparison sample's firms among those the tests load, from every construction of patent technologies and peers, on the observed assignment and on every draw, changes no decision.

The order-and-timing design. The unit is the cell, a firm and a CPC main group with open-source activity and a linked patent whose first filing is 2016 or later, so that code could have been seen first: 3,330 cells of 599 firms, on corrected event dates. The reference permutes, across the firm's own repositories, when each repository's activity happens, keeping every repository's labels and the whole patent side, so it breaks only which of the firm's technologies saw code when; Holm

runs across the three statistics of Table B18, declared before the first output, and Table B19 gives the re-run on recent repositories, the drop of shared patents and the dating of patents by grant. Henkel (2006) gives a mechanism only, what is revealed rather than when, and von Hippel and von Krogh (2003) the setting.

Table B17: The same split on the firm's patents against the object-level reference (extends Table B16)

Statistic	Observed	Reference mean	Holm p
Dispersion across firms of patent-side outside rates	0.073	0.0932	0.0299
Dispersion across technologies	0.0072	0.0458	—
Agreement of the two channels' rankings	0.2312	0.3256	—
Mean patent-side outside rate	0.8602	0.6681	
Dispersion across firms, scaled by the rate's ceiling	0.6068	0.4203	
Dispersion across technologies, scaled by the rate's ceiling	0.0601	0.2065	

The split of Table B16 run on the firm's patents: on 181,890 patent-active cell rows of 586 firms, 200 draws. The raw dispersions mostly restate a level difference: real open-source mass concentrates in 528 technologies, while a draw spreads it over about 2,524. The scaled rows divide by each assignment's rate ceiling. It is descriptive: no reason for opening signs its statistics without a stated configuration.

Table B18: Order and timing in the cells where a firm has both channels

Statistic	Observed	Two-sided p	Holm p
Code comes first, share of cells	41.56%	0.004 [0.002]	0.012 [0.006]
Median gap in years, filing minus code	0.0	1.0 [1.0]	—
Share of code activity in the two years ending at the filing	13.32%	0.008 [0.006]	0.016 [0.012]
Cells	3,330, of 599 firms		

A cell is a firm and a CPC main group where the firm has open-source activity and a linked patent, with a first filing in 2016 or later. The reference permutes, across the firm's own repositories, when each repository's activity happens; p is the recorded two-sided rank p-value against 1,000 reference rows, which include the observed assignment and count it among the draws as well as in the plus-one; in brackets, the companion by the plus-one rule over the permutations alone (Phipson and Smyth, 2010). Holm runs across the three statistics. Corrected event dates. The re-run on recent repositories, the drop of shared patents and the dating of patents by grant year are in Table B19.

B.8 What moves with the distances

What moves with a firm's distances is mostly their references, as Lemma 8 leads one to expect: each distance's excess over its reference rises with the firm's count in the channel the distance starts from, while the observed distance moves less or falls, on the classes that carry code as on all of the firm's patents. Every regression with a distance puts the distance, or its excess over a named reference, on the left and reads each coefficient against the same regression on the reference. The distance's level under random assignment falls with the counts that build it (Lemma 8), so the counts go on the right; and a firm characteristic has no general sign for it, so every comparison is read two-sided. Every column absorbs its own fixed effects, clusters its standard errors by firm and runs on one sample. The count slope is one of four comparisons fixed in advance, the other three, on R&D intensity, size and industry, in Appendix B.11; every other table is a description. The

Table B19: Order and timing on recent repositories and with patents dated by grant (extends Table B18)

	Primary	Code first observed from 2020	Patents dated by grant
Code comes first, share of cells	41.56%	4.83%	53.09%
two-sided p	0.004 [0.002]	0.004 [0.002]	0.004 [0.002]
Median gap in years	0.0	−3.0	1.0
two-sided p	1.0 [1.0]	0.004 [0.002]	— [1.0]
Share of code activity around the filing	13.32%	4.2%	17.7%
two-sided p	0.008 [0.006]	— [0.002]	0.37 [0.368]
Cells	3,330	2,549	6,321

The companion keeps the repositories first observed from 2020, whose labels are closest to their content's date; one p is recorded for its first two statistics. In brackets, the companion p by the plus-one rule over the permutations alone (Table B18); where no p is recorded, the companion stands alone. The last column dates a cell's patents by their first grant year, and the around-filing share becomes the share around the grant. Dropping the 16,487 patents attached to two or more of the panel's parents leaves the 3,330 cells and every statistic unchanged; its recorded p for the around-filing share is 0.016, and its companions are 0.002, 1.0 and 0.014 for the three statistics.

primary family of distance regressions is withdrawn by a dated amendment to its specification, so no member is read as a test (Appendix B.9).

B.8.1 The distance and the scale of open-source activity

The firm's excess over the randomization benchmark rises with the scale of its open-source activity, the log of one plus its repository count, mostly because the randomization benchmark's mean falls as the count grows, and it does so on the classes that carry code as on all of the firm's patents.

The count slope. The proposition tested is the paper's own: by Lemma 8, the distance's level under random assignment falls with the counts that build it, so a firm with more repositories sits further above its reference even where its code does not move. The count slope matches it. Firm-years with more repositories are more separated than the randomization benchmark: with firm and year effects, on the comparison sample, the slope's excess over the randomization benchmark is 0.0516 (standard error 0.0072), and the fixed-portfolio world reproduces it, at 0.0463. The slope is β_{repo} in

$$D_{it} = \beta_{\text{repo}} \log(1 + n^{\text{repo}}_{i,t-1}) + \beta_{\text{ev}} \log(1 + n^{\text{ev}}_{i,t-1}/n^{\text{repo}}_{i,t-1}) + \beta_{\text{pat}} \log(1 + n^{\text{pat}}_{i,t-1}) + \alpha_i + \lambda_t + \varepsilon_{it}, \quad (4)$$

where D_{it} is the distance under event weights over the five-year window that ends in year t, and $n^{\text{repo}}_{i,t-1}$, $n^{\text{ev}}_{i,t-1}$ and $n^{\text{pat}}_{i,t-1}$ are the firm's repositories, open-source events and patents over the window that ends in $t-1$; $\log(1 + n^{\text{repo}}_{i,t-1})$ measures the scale of the firm's open-source activity, not its intent. The two windows share four years, so the slope is an association between overlapping histories, not a response to earlier activity. Errors are clustered by parent firm (Cameron and Miller, 2015, §I); and since the distance lies between zero and one, with 6.35% of firm-years at one at main group, the linear slope is read as an average association (Papke and Wooldridge, 1996, p. 619). The same regression on the excess over the randomization benchmark's mean gives the slope's excess. Appendix B.1 shows that within firms the slope mostly reflects how many units are drawn, that about half of it survives across firms at a common depth, that the content of firms with similar

Table B20: Which patenting firms start publishing code: the full hazard models (extends Table E4)

A. The four own-class forms

	Own classes	Own classes; grant stock, industry	With the norm	With the norm; grant stock, industry
Software-heavy patent stock	5.035	3.068	4.469	2.830
	[3.599, 7.043]	—	[3.141, 6.358]	—
Technology norm, per SD			1.136 (p 0.0331)	1.103 (p 0.168)
Dominant core business	0.973	0.982	0.894	0.960
	[0.742, 1.274]	—	[0.672, 1.191]	—
Size (log assets), per SD	1.444	1.767	1.455	1.756
Proportional-hazards test, p	0.0248	0.4280	0.0136	0.0601
Spells / first entries	3,382 / 214	—	3,381 / 214	—

B. Software-heavy patent stock under other event dates and rules

	Own classes	With the norm	Spells	First entries
Dated by creation year, qualifying license	5.734 [3.513, 9.357]	4.973	3,364	106
Dated by creation year, any license	5.154 [3.564, 7.454]	4.683	3,358	165
Shared patents dropped, covariates rebuilt	5.068 [3.609, 7.117]	4.299 [3.002, 6.157]	3,170	210
and eligibility rebuilt	5.191 [3.697, 7.287]	4.475 [3.128, 6.403]	3,236	211

C. The fading, and firm frailty

	Hazard ratio	95% interval	Spells or records	First entries
Software-heavy at 2020, time-varying form	5.106	[3.593, 7.257]	3,382	214
per-year multiplier (p)	0.843 (0.037)			
Software-heavy, 2016–2019	8.146	[3.672, 18.071]	862	65
Software-heavy, 2020–2024	2.879	[1.337, 6.199]	789	38
Stacked windows, Cox: software-heavy, early window	6.417	[3.831, 10.748]	5,706	214
software × late window (p)	0.654 (0.225)	[0.329, 1.298]		
Stacked windows, Weibull: software-heavy, early window	6.324	[3.773, 10.600]	5,706	214
software × late window (p)	0.671 (0.256)	[0.338, 1.335]		
Stacked windows, Weibull with gamma frailty: software-heavy, early window	12.159	[6.134, 24.101]	5,706	214
software × late window (p)	2.559 (0.054)	[0.983, 6.660]		
Frailty variance θ	18.177			

D. The covariates of the other reasons for opening, in one model

	Hazard ratio	95% interval	Raw p	Holm p
Gross margin, per rank SD	1.19	[0.94, 1.51]	0.139	0.416
R&D intensity, per rank SD	1.01	[0.74, 1.38]	0.935	0.935
Software type (industry rules)	2.49	[1.11, 5.60]	0.027	0.109
Rivals' earlier entry, per SD	0.88	[0.71, 1.10]	0.262	0.525
Software-heavy patent stock	4.43	[2.43, 8.09]		
Size (log assets), per SD	1.86	[1.27, 2.75]		
No reported R&D	0.35	—	0.00097	

Cox hazard on calendar years, Efron ties, robust standard errors, unless the row says otherwise; 95% intervals in brackets. The proportional-hazards test is that of Grambsch and Therneau (1994, §2, eqs. 9–11, p. 517). Panel A's first and third columns are Table E4. Panel B: dating by creation year re-times each entry by the repository's creation year, with or without the rule that the repository carry a recognized license; the crosswalk check drops the patents attached to two or more firms of the risk set. Panel C: the time-varying form interacts the covariates with the year, centered at 2020; the windows are calendar windows; the stacked forms hold one record per firm and window, and the frailty form adds a firm-level gamma frailty to a Weibull baseline. Panel D is the model that carries all four contrasts, on 984 spells with 103 first entries; its proportional-hazards test has p 0.136.

patent portfolios reproduces it, and how it varies with the window, the industry and the classes that carry code (Table B2).

The excess and the observed distance. Table 7 describes the same relation with more of the firm's variables and the fixed effects switched by column, with the excess over the randomization benchmark's mean, $\tilde{D}_{it}$, and the observed distance, D_{it}, on the left, on the same firm-years and the same right-hand side. By the least-squares identity, each coefficient on the excess equals the coefficient on the observed distance minus the coefficient on the randomization benchmark's mean.

Table B21: The firm's own distance and its counts

	(1)	(2)	(3)	(4)	(5)	(6)
A. The excess over the randomization benchmark, $\tilde{D}_{it}$						
Log repositories $(t-1)$	0.016	0.017	0.017	0.017	0.031	0.032
	(0.008)	(0.009)	(0.008)	(0.009)	(0.013)	(0.013)
Log events per repository $(t-1)$	−0.02	−0.019	−0.02	−0.02	−0.008	−0.008
	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.008)
Log patents $(t-1)$	0.01	0.009	0.024	0.024	0.02	0.021
	(0.006)	(0.006)	(0.006)	(0.007)	(0.01)	(0.01)
Log assets $(t-1)$	−0.001	−0.00038	−0.01	−0.01	0.003	0.004
	(0.007)	(0.007)	(0.009)	(0.01)	(0.01)	(0.01)
B. The observed distance, D_{it}, on the same right-hand side						
Log repositories $(t-1)$	−0.054	−0.054	−0.035	−0.035	−0.014	−0.014
	(0.008)	(0.008)	(0.009)	(0.01)	(0.008)	(0.008)
Log events per repository $(t-1)$	0.012	0.013	0.019	0.02	0.002	0.002
	(0.008)	(0.008)	(0.008)	(0.009)	(0.005)	(0.005)
Log patents $(t-1)$	0.006	0.005	0.00044	−0.000088	0.017	0.017
	(0.007)	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)
Log assets $(t-1)$	0.02	0.02	0.014	0.014	0.010	0.01
	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)
Firm					Yes	Yes
Year		Yes				Yes
Industry			Yes			
Industry × year				Yes		
Firm-years	883	883	883	883	883	883
Firms	226	226	226	226	226	226

Each column regresses the firm-year's own distance under event weights, or its excess over the randomization benchmark's mean, on the firm's variables over the five-year window ending $t-1$; both panels use the same firm-years and the same right-hand side, which also carries R&D intensity, the leader indicator among US-dollar reporters, panel age, the exposure to the 2019 eligibility guidance, product-market closeness to the industry and first-time outside participants (Appendix Table B27). Standard errors clustered by firm, in parentheses. Industry: NAICS four digits, three where fewer than five firms have linked code. A description, not a test.

What moves with the firm's counts is mostly the reference. Within a firm (Table 7, column 6, firm and year effects), the excess rises with the repository count, 0.032 (0.013), while the observed distance moves within two standard errors of zero, −0.014 (0.008). Across firms (columns 1 to 4) the observed distance falls with the repository count and the excess rises, the same identity between firms. With the counts and size alone, on the firm-years with size in US dollars (Appendix Table B26, panels A and B), the excess rises with the repository count within firms, 0.046 (0.009), and the observed distance moves with it by less, −0.016 (0.005), within two standard errors of zero. Run the same way with the randomization benchmark's mean on the left, on the same firm-years, the reference moves with the repository count by −0.0613 (0.0071), and the three regressions add

up term by term, the excess equal to the observed minus the reference up to rounding.

On the classes that carry code. The same proposition holds when the firm's patents are restricted to the classes that carry code: Lemma 8 leads one to expect the excess there to rise with the repository count, and it does. The *GitHub classes* GH are the 723 CPC main groups that carry open-source mass in any firm-year window, and the *non-GitHub share* ψ_{it} is the share of the firm-year's patent mass outside them; D_{GH} is the distance between the firm's patents on GH, rescaled to sum to one, and its code, all of which lies in GH (Appendix B.4). Without firm effects, with two-digit NAICS and year effects and ψ_{it} as a control, the slope of D_{GH} on the repository count is −0.0399 (0.0044) against the randomization benchmark's −0.0745, so its excess over the randomization benchmark rises by 0.0345 (0.0044) (Table B8); on the same rows the full distance's slope is −0.0393 (0.0043), and Tables B22 and B23 give every coefficient of both. On GH, code is the more concentrated channel in 84% of firm-years, mostly as the randomization benchmark makes it, and code and patents share a main group in 95.3% of firm-years against 99.8% under the randomization benchmark (Appendix B.4).

B.8.2 The distances to the industry

The proposition tested is again Lemma 8's: same-size draws from the industry's pools sit nearer the pools as the firm's portfolio grows, so each distance's excess over its declared reference should rise with the firm's count in the channel the distance starts from, and the observed distance fall. The four distances from the firm to its industry match it; the industry's own distance, $D(P_{G,t}, O_{G,t})$, starts from no channel of the firm's, so the lemma makes no prediction for it, and it is described. For each of the five distances that involve the firm's industry (Appendix H), with $\tilde{D}(\cdot,\cdot)$ its excess over its declared reference (the industry-pool benchmark for the four distances from the firm to its industry, the within-industry benchmark for $D(P_{G,t}, O_{G,t})$),

$$\tilde{D}(\cdot,\cdot) = \boldsymbol{\beta}^{\top}\mathbf{x}_{i,t-1} + \phi_{it} + \varepsilon_{it}, \tag{5}$$

and the same with the observed distance on the left, where $\mathbf{x}_{i,t-1}$ is the short list: the three counts, log assets in US dollars, R&D intensity and panel age, and ϕ_{it} the fixed effects each column absorbs.

Within a firm (Table B25, panel A), the firm's code to its industry's code rises against the industry-pool benchmark with the repository count, 0.011 (0.009), within two standard errors of zero, and falls observed, −0.047 (0.011); the firm's patents to its industry's patents rise against the industry-pool benchmark with the patent count, 0.055 (0.004), and fall observed, −0.035 (0.006). The firm's code to its industry's patents and its patents to its industry's code follow the same pattern, and so does each across firms within the industry-year (panel B). Run the same way with the industry-pool benchmark's mean on the left, the reference falls with the firm's count faster than the observed distance does: within a firm, at −0.0579 (0.006) with the repository count for the firm's code to its industry's code, and at −0.0902 (0.0058) with the patent count for its patents to its

Table B22: The distance within the GitHub classes, D_{GH}, on the count regressors and ψ_{it}: every coefficient (extends Table B8)

Fixed effects	None	Year	NAICS-2, year	NAICS-3, year
A. D_{GH}, event weights				
$\log(1+n^{\text{repo}})$	−0.0416	−0.0417	−0.0399	−0.0385
	(0.0042)	(0.0042)	(0.0044)	(0.0044)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	0.0008	0.0006	0.0009	0.0046
	(0.004)	(0.004)	(0.0041)	(0.0042)
$\log(1+n^{\text{pat}})$	0.0027	0.0027	0.0012	−0.0021
	(0.0027)	(0.0027)	(0.0029)	(0.0029)
ψ_{it}	0.2476	0.2467	0.2206	0.2374
	(0.0174)	(0.0175)	(0.0229)	(0.0257)
Constant	0.8911	0.8924	0.8957	0.8844
	(0.023)	(0.0231)	(0.0237)	(0.0243)
Firm-years	2,403	2,403	2,403	2,403
Parent firms	579	579	579	579
R^2	0.352	0.353	0.378	0.419
Within R^2	0.352	0.353	0.236	0.204
B. D_{GH}, equal weights				
$\log(1+n^{\text{repo}})$	−0.0422	−0.0423	−0.0408	−0.0385
	(0.0042)	(0.0042)	(0.0044)	(0.0044)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0047	−0.0047	−0.0035	−0.0004
	(0.004)	(0.004)	(0.004)	(0.0041)
$\log(1+n^{\text{pat}})$	−0.0033	−0.0033	−0.0055	−0.0094
	(0.0026)	(0.0026)	(0.0029)	(0.0028)
ψ_{it}	0.3012	0.3008	0.2589	0.2541
	(0.0178)	(0.0179)	(0.0219)	(0.0245)
Constant	0.8962	0.8967	0.9025	0.8971
	(0.0229)	(0.023)	(0.0235)	(0.0234)
Firm-years	2,403	2,403	2,403	2,403
Parent firms	579	579	579	579
R^2	0.439	0.439	0.472	0.514
Within R^2	0.439	0.439	0.312	0.271

n^{repo}, n^{ev} and n^{pat} are repositories, open-source events and patents over the five-year window that ends in $t-1$; ψ_{it} is from the outcome's own window. The rows are the firm-years with a two-digit NAICS code and $\psi_{it} < 1$; Appendix Table B10 adds the columns with size and R&D, on their own firm-years, and with firm and year effects. Standard errors clustered by parent firm in parentheses.

Table B23: The full distance on the count regressors and ψ_{it}, on the same rows (extends Table B8)

Fixed effects	None	Year	NAICS-2, year	NAICS-3, year
A. D_{it}, event weights				
$\log(1+n^{\text{repo}})$	−0.0413	−0.0415	−0.0393	−0.0379
	(0.0041)	(0.0041)	(0.0043)	(0.0043)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0008	−0.0009	−0.0001	0.0036
	(0.0039)	(0.0039)	(0.004)	(0.0041)
$\log(1+n^{\text{pat}})$	0.0041	0.004	0.002	−0.0009
	(0.0026)	(0.0026)	(0.0027)	(0.0028)
ψ_{it}	0.2819	0.281	0.2516	0.2697
	(0.0171)	(0.0172)	(0.0219)	(0.0245)
Constant	0.8972	0.8985	0.9007	0.8876
	(0.022)	(0.0221)	(0.0226)	(0.0235)
Firm-years	2,403	2,403	2,403	2,403
Parent firms	579	579	579	579
R^2	0.404	0.405	0.428	0.463
Within R^2	0.404	0.404	0.266	0.227
B. Equal weights				
$\log(1+n^{\text{repo}})$	−0.0418	−0.0418	−0.0399	−0.0378
	(0.0041)	(0.0041)	(0.0043)	(0.0043)
$\log(1+n^{\text{ev}}/n^{\text{repo}})$	−0.0063	−0.0064	−0.0046	−0.0015
	(0.0039)	(0.0039)	(0.0039)	(0.0039)
$\log(1+n^{\text{pat}})$	−0.001	−0.001	−0.0037	−0.0074
	(0.0025)	(0.0025)	(0.0027)	(0.0027)
ψ_{it}	0.347	0.3466	0.3047	0.3052
	(0.0177)	(0.0178)	(0.0215)	(0.0236)
Constant	0.8985	0.899	0.9026	0.8959
	(0.0219)	(0.022)	(0.0225)	(0.0227)
Firm-years	2,403	2,403	2,403	2,403
Parent firms	579	579	579	579
R^2	0.5	0.5	0.529	0.567
Within R^2	0.5	0.5	0.353	0.302

n^{repo}, n^{ev} and n^{pat} are repositories, open-source events and patents over the five-year window that ends in $t-1$; ψ_{it} is from the outcome's own window. The rows are the firm-years with a two-digit NAICS code and $\psi_{it} < 1$; Appendix Table B11 adds the columns with size and R&D, on their own firm-years, and with firm and year effects. Standard errors clustered by parent firm in parentheses.

Table B24: The excess by the size of the open-source program

Repositories in the window	Firm-years	Excess (SE)	Share outside (SE)
Bottom quartile	617	0.1855 (0.0091)	1.4693 (0.0992)
Second quartile	619	0.2179 (0.0102)	1.157 (0.0925)
Third quartile	617	0.2545 (0.0101)	0.7517 (0.0706)
Top quartile	619	0.2762 (0.0098)	0.3056 (0.0487)
Top decile	250	0.2864 (0.0152)	0.1045 (0.0446)

Event weights; the comparison sample by quartile of the firm-year's repository count, and its top decile. The excess is the observed distance minus the randomization benchmark's mean; the share outside is the part of it in cells outside the firm's patent technologies, unweighted. A share above one means the inside part's excess is negative: code inside the patent technologies sits nearer the patents than the randomization benchmark places it. Standard errors clustered by firm. Equality across the quartiles: F 17.01 under event weights and 29.54 under equal weights.

Table B25: The distances to the industry and the firm's variables

	$D(O_{it}, O_{G,t})$		$D(O_{it}, P_{G,t})$		$D(P_{it}, O_{G,t})$		$D(P_{it}, P_{G,t})$		$D(P_{G,t}, O_{G,t})$	
	$\tilde{D}$	D	$\tilde{D}$	D	$\tilde{D}$	D	$\tilde{D}$	D	$\tilde{D}$	D
A. Within the firm: firm and year effects										
Log repositories $(t-1)$	0.011	−0.047	0.012	−0.016	0.005	−0.003	−0.001	−0.005	0.006	−0.003
	(0.009)	(0.011)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.002)
Log events per repository $(t-1)$	−0.012	0.001	−0.003	0.005	−0.004	−0.006	−0.004	−0.004	0.004	−0.003
	(0.007)	(0.007)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)
Log patents $(t-1)$	−0.004	0.01	0.001	0.002	0.01	−0.014	0.055	−0.035	−0.012	−0.001
	(0.008)	(0.008)	(0.003)	(0.003)	(0.005)	(0.005)	(0.004)	(0.006)	(0.004)	(0.001)
Log assets $(t-1)$	−0.007	−0.008	−0.000060	0.002	−0.006	−0.01	0.003	−0.007	0.001	−0.001
	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)	(0.002)
R&D intensity $(t-1)$	0.063	0.057	0.002	0.003	−0.026	−0.037	0.008	−0.017	0.02	−0.002
	(0.054)	(0.049)	(0.013)	(0.012)	(0.018)	(0.023)	(0.017)	(0.018)	(0.016)	(0.005)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
B. Across firms within the industry-year: industry × year effects										
Log repositories $(t-1)$	0.026	−0.046	0.015	−0.03	−0.019	−0.02	−0.013	−0.004	0.00023	0.00049
	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)	(0.006)	(0.006)	(0.002)	(0.001)
Log events per repository $(t-1)$	−0.02	0.002	−0.007	0.004	−0.002	−0.007	0.005	−0.003	0.003	−0.002
	(0.007)	(0.007)	(0.003)	(0.003)	(0.005)	(0.005)	(0.006)	(0.006)	(0.001)	(0.001)
Log patents $(t-1)$	0.002	0.00018	0.002	−0.000055	0.012	−0.01	0.045	−0.049	−0.00011	0.001
	(0.006)	(0.006)	(0.002)	(0.003)	(0.005)	(0.005)	(0.007)	(0.006)	(0.002)	(0.002)
Log assets $(t-1)$	−0.012	−0.015	0.001	0.002	−0.013	−0.011	−0.015	−0.017	−0.001	0.001
	(0.007)	(0.008)	(0.003)	(0.004)	(0.005)	(0.006)	(0.008)	(0.007)	(0.002)	(0.002)
R&D intensity $(t-1)$	0.03	0.025	0.028	0.038	−0.112	−0.161	0.041	−0.053	0.008	0.009
	(0.079)	(0.084)	(0.035)	(0.034)	(0.073)	(0.076)	(0.064)	(0.057)	(0.013)	(0.013)
Panel age	−0.003	−0.001	−0.003	−0.002	0.001	−0.001	0.00024	0.00017	−0.00032	0.00017
	(0.003)	(0.003)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.001)	(0.001)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356

Each column regresses one distance under event weights (D) or its excess over its declared reference ($\tilde{D}$: the industry-pool benchmark for the four distances from the firm to its industry, the within-industry benchmark for the industry's own two portfolios) on the firm's variables: the three counts over the five-year window ending $t-1$, log assets in US dollars, R&D intensity and panel age (the short list). The industry is the firm's NAICS group in the year before, the firm left out. Panel A absorbs firm and year effects (panel age is absorbed, a dash); panel B absorbs industry-by-year cells. Standard errors clustered by firm, in parentheses. Every column of the ladder, and the full list of variables, is in Appendix B.9. A description, not a test.

industry's patents. This is the small-unit bias of section F.7, read firm by firm. The industry's own two portfolios, with the firm left out, move little with the firm's variables within the industry-year, where they vary only through which firm is left out. Every column, with the full list of variables, is in Appendix B.9.

B.9 The distances on the left: every column

The tables below extend those of Appendix B.8. Each is a ladder of fixed effects on one sample: the columns switch the absorbed effects on and off on the same firm-years, as the bottom rows mark. Forms: the excess of a distance from its declared reference, and the observed distance on the same right-hand side. Lists: the short list on the firm-years with size in US dollars, and the full list on the firm-years where every variable is defined. Size is log assets where the firm's Compustat North America row for the year before reports in US dollars, 2,258 of the 3,064 firm-years with assets, and the leader indicator and the R&D rank are ranked among US-dollar reporters. Every table is a description: the declared primary family of distance regressions is withdrawn by a dated amendment to its specification, and the leader indicator and the exposure to the 2019 guidance, which its members use, enter the full lists here only as descriptions. The family has four members, position against peers' code, position against peers' patents, same-industry peers' patents where the firm opened, and the 2019 guidance, with Holm's adjustment across the four. The amendment withdraws it because its adjustment needs four p-values and three cannot be had: two members are read against a reference that holds an industry norm, which is not built, one waits for peers rebuilt on corrected dates, and one has no form with a reference difference.

B.10 The excess under other readings of the classification

In every reading of the classification, the excess over the randomization benchmark has its 95 percent interval above zero. Table B40 reads both channels in other ways: at the subclass and the class level of the CPC, without the software-engineering main group, with the software main groups G06F 8/00 to 11/00 merged, and with each item in a single main group.

B.11 The comparisons declared before testing

Each of the four comparisons declared before testing asks whether a variable goes with the firm's distance by more than it does under the randomization benchmark. This appendix gives the method, the equation, the definitions, Holm's adjustment and the results of the four comparisons that section 5.3 reports: the count of open-source projects, R&D intensity, size and industry (Table B41). The first tests the count slope, and the other three ask whether size, R&D intensity and industry go with the excess across firms. The benchmark distance itself moves with the counts, and it can move with the comparison's variable too, so the observed coefficient alone mixes what the randomization benchmark produces with what goes with the excess. I therefore run the same regression on each draw of the randomization benchmark, so that each draw gives the coefficient

Table B26: The firm's own distance on the short list: every column (extends Table 7)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Excess over the randomization benchmark							
Log repositories	0.027	0.027	0.028	0.028	0.031	0.046	0.042
	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.009)	(0.009)
Log events per repository	−0.02	−0.02	−0.019	−0.019	−0.014	−0.012	−0.013
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Log patents	0.005	0.004	0.009	0.009	0.007	0.012	0.014
	(0.004)	(0.004)	(0.004)	(0.005)	(0.007)	(0.008)	(0.009)
Log assets	0.002	0.002	−0.00043	−0.000081	−0.009	−0.002	−0.001
	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Firm-years	1,703	1,703	1,703	1,703	1,703	1,703	1,703
Firms	412	412	412	412	412	412	412
B. Observed							
Log repositories	−0.066	−0.066	−0.052	−0.052	−0.017	−0.016	−0.016
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Log events per repository	−0.007	−0.008	0.001	0.001	0.003	0.003	0.005
	(0.005)	(0.005)	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)
Log patents	0.007	0.007	−0.005	−0.005	0.004	0.005	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)
Log assets	0.017	0.017	0.018	0.018	0.001	0.001	0.005
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Firm-years	1,703	1,703	1,703	1,703	1,703	1,703	1,703
Firms	412	412	412	412	412	412	412
Firm					Yes	Yes	Yes
Year		Yes				Yes	
Industry			Yes				
Industry × year				Yes			Yes

Event weights. Panel A: the excess over the randomization benchmark's mean; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where fewer than five firms have linked code.

Table B27: The firm's own distance on the full list, the common sample: every column (extends Table 7)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Excess over the randomization benchmark							
Log repositories	0.016	0.017	0.017	0.017	0.031	0.032	0.026
	(0.008)	(0.009)	(0.008)	(0.009)	(0.013)	(0.013)	(0.014)
Log events per repository	−0.02	−0.019	−0.02	−0.02	−0.008	−0.008	−0.009
	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)
Log patents	0.01	0.009	0.024	0.024	0.02	0.021	0.022
	(0.006)	(0.006)	(0.006)	(0.007)	(0.01)	(0.01)	(0.009)
Log assets	−0.001	−0.00038	−0.01	−0.01	0.003	0.004	−0.002
	(0.007)	(0.007)	(0.009)	(0.01)	(0.01)	(0.01)	(0.011)
R&D intensity	0.058	0.067	−0.032	−0.025	−0.005	−0.009	−0.01
	(0.092)	(0.094)	(0.073)	(0.082)	(0.023)	(0.023)	(0.021)
Panel age	0.002	0.003	0.003	0.004	−0.006	—	—
	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)		
Leader (top third, R&D rank)	−0.006	−0.006	−0.016	−0.02	0.001	−0.000070	−0.00018
	(0.021)	(0.021)	(0.021)	(0.022)	(0.02)	(0.02)	(0.02)
Product-market closeness $(t-1)$	0.431	0.424	0.305	0.369	−0.362	−0.376	−0.436
	(0.505)	(0.508)	(0.5)	(0.545)	(0.214)	(0.221)	(0.241)
Exposure to the 2019 guidance	−0.0000060	−0.0000062	−0.000016	−0.000018	—	—	—
	(0.000016)	(0.000016)	(0.000019)	(0.000019)			
Log first-time participants	0.004	0.003	0.003	0.003	−0.00012	−0.001	0.001
	(0.007)	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)
Firm-years	883	883	883	883	883	883	883
Firms	226	226	226	226	226	226	226
B. Observed							
Log repositories	−0.054	−0.054	−0.035	−0.035	−0.014	−0.014	−0.016
	(0.008)	(0.008)	(0.009)	(0.01)	(0.008)	(0.008)	(0.008)
Log events per repository	0.012	0.013	0.019	0.02	0.002	0.002	0.003
	(0.008)	(0.008)	(0.008)	(0.009)	(0.005)	(0.005)	(0.005)
Log patents	0.006	0.005	0.00044	−0.000088	0.017	0.017	0.013
	(0.007)	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)	(0.01)
Log assets	0.02	0.02	0.014	0.014	0.010	0.01	0.01
	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)	(0.009)
R&D intensity	−0.017	−0.005	−0.078	−0.079	−0.015	−0.015	−0.014
	(0.093)	(0.095)	(0.072)	(0.076)	(0.017)	(0.017)	(0.019)
Panel age	0.007	0.008	0.005	0.006	−0.002	—	—
	(0.003)	(0.004)	(0.003)	(0.004)	(0.002)		
Leader (top third, R&D rank)	−0.046	−0.047	−0.047	−0.048	−0.011	−0.012	−0.013
	(0.02)	(0.02)	(0.022)	(0.023)	(0.016)	(0.015)	(0.016)
Product-market closeness $(t-1)$	0.003	−0.002	−0.344	−0.302	−0.283	−0.292	−0.304
	(0.414)	(0.418)	(0.401)	(0.435)	(0.159)	(0.165)	(0.188)
Exposure to the 2019 guidance	0.000032	0.000032	0.000040	0.000039	—	—	—
	(0.000024)	(0.000024)	(0.000027)	(0.000028)			
Log first-time participants	−0.021	−0.021	−0.024	−0.025	−0.01	−0.01	−0.008
	(0.007)	(0.007)	(0.007)	(0.008)	(0.005)	(0.005)	(0.004)
Firm-years	883	883	883	883	883	883	883
Firms	226	226	226	226	226	226	226
Firm					Yes	Yes	Yes
Year		Yes				Yes	
Industry			Yes				
Industry × year				Yes			Yes

Event weights. Panel A: the excess over the randomization benchmark's mean; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where fewer than five firms have linked code.

Table B28: The firm's code and the industry's code, $D(O_{it}, O_{G,t})$, on the short list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.02	0.02	0.011	0.011	0.026	0.024	0.026	0.024
	(0.005)	(0.005)	(0.009)	(0.009)	(0.006)	(0.006)	(0.006)	(0.006)
Log events per repository	−0.022	−0.022	−0.012	−0.012	−0.02	−0.019	−0.02	−0.019
	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
Log patents	0.008	0.008	−0.004	−0.004	0.001	0.001	0.002	0.001
	(0.005)	(0.006)	(0.008)	(0.008)	(0.006)	(0.006)	(0.006)	(0.006)
Log assets	−0.013	−0.013	−0.007	−0.007	−0.012	−0.009	−0.012	−0.009
	(0.006)	(0.006)	(0.008)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)
R&D intensity	0.032	0.031	0.063	0.063	0.029	0.028	0.03	0.028
	(0.066)	(0.067)	(0.054)	(0.054)	(0.075)	(0.076)	(0.079)	(0.079)
Panel age	−0.002	−0.002	0.001	—	−0.003	−0.003	−0.003	−0.003
	(0.003)	(0.003)	(0.002)		(0.003)	(0.003)	(0.003)	(0.003)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
B. Observed								
Log repositories	−0.061	−0.061	−0.046	−0.047	−0.047	−0.048	−0.046	−0.047
	(0.006)	(0.006)	(0.01)	(0.011)	(0.006)	(0.006)	(0.006)	(0.006)
Log events per repository	−0.004	−0.005	0.001	0.001	0.002	0.006	0.002	0.006
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Log patents	0.016	0.017	0.01	0.01	0.00039	−0.002	0.00018	−0.002
	(0.006)	(0.006)	(0.008)	(0.008)	(0.006)	(0.006)	(0.006)	(0.006)
Log assets	−0.018	−0.018	−0.008	−0.008	−0.015	−0.014	−0.015	−0.014
	(0.007)	(0.007)	(0.009)	(0.009)	(0.007)	(0.007)	(0.008)	(0.008)
R&D intensity	0.022	0.011	0.054	0.057	0.024	0.031	0.025	0.031
	(0.069)	(0.071)	(0.048)	(0.049)	(0.079)	(0.079)	(0.084)	(0.082)
Panel age	0.003	0.002	0.005	—	−0.001	−0.001	−0.001	−0.001
	(0.003)	(0.003)	(0.003)		(0.003)	(0.003)	(0.003)	(0.003)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B29: The firm's code and the industry's code, $D(O_{it}, O_{G,t})$, on the full list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.016	0.015	0.004	0.004	0.022	0.024	0.021	0.024
	(0.008)	(0.009)	(0.01)	(0.01)	(0.009)	(0.009)	(0.009)	(0.009)
Log events per repository	−0.032	−0.033	−0.013	−0.013	−0.025	−0.025	−0.027	−0.026
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.01)	(0.01)
Log patents	0.01	0.011	0.007	0.007	0.002	0.001	0.002	0.001
	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.01)
Log assets	−0.016	−0.017	−0.006	−0.007	−0.01	−0.003	−0.01	−0.003
	(0.009)	(0.009)	(0.011)	(0.012)	(0.009)	(0.01)	(0.01)	(0.01)
R&D intensity	0.024	0.015	0.014	0.014	0.057	0.081	0.059	0.082
	(0.084)	(0.085)	(0.023)	(0.023)	(0.084)	(0.082)	(0.09)	(0.086)
Panel age	−0.002	−0.003	0.004	—	−0.004	−0.002	−0.004	−0.002
	(0.004)	(0.004)	(0.003)		(0.004)	(0.004)	(0.004)	(0.004)
Leader (top third, R&D rank)	0.005	0.006	0.003	0.004	−0.007	−0.034	−0.007	−0.034
	(0.023)	(0.023)	(0.01)	(0.01)	(0.026)	(0.025)	(0.028)	(0.026)
Product-market closeness ($t-1$)	1.061	1.066	0.308	0.32	0.933	1.272	0.974	1.305
	(0.476)	(0.481)	(0.277)	(0.282)	(0.456)	(0.534)	(0.488)	(0.559)
Exposure to the 2019 guidance	−0.000055	−0.000055	—	—	−0.000037	−0.000057	−0.000038	−0.000057
	(0.000018)	(0.000018)			(0.000019)	(0.000021)	(0.000020)	(0.000021)
Log first-time participants	0.014	0.015	0.003	0.003	0.011	0.011	0.011	0.011
	(0.007)	(0.007)	(0.005)	(0.005)	(0.007)	(0.007)	(0.008)	(0.007)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
B. Observed								
Log repositories	−0.065	−0.067	−0.043	−0.044	−0.046	−0.044	−0.046	−0.044
	(0.01)	(0.01)	(0.013)	(0.013)	(0.01)	(0.009)	(0.011)	(0.01)
Log events per repository	−0.009	−0.011	−0.001	−0.001	0.004	0.005	0.003	0.005
	(0.01)	(0.01)	(0.009)	(0.009)	(0.01)	(0.009)	(0.011)	(0.009)
Log patents	0.024	0.026	0.029	0.028	0.002	−0.004	0.002	−0.004
	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)
Log assets	−0.029	−0.03	−0.003	−0.004	−0.016	−0.008	−0.016	−0.008
	(0.009)	(0.009)	(0.013)	(0.014)	(0.01)	(0.01)	(0.011)	(0.01)
R&D intensity	−0.03	−0.05	0.027	0.032	−0.001	0.055	−0.004	0.053
	(0.088)	(0.091)	(0.026)	(0.028)	(0.096)	(0.087)	(0.106)	(0.093)
Panel age	0.004	0.002	0.006	—	−0.001	0.001	−0.001	0.001
	(0.004)	(0.004)	(0.004)		(0.004)	(0.004)	(0.004)	(0.004)
Leader (top third, R&D rank)	0.001	0.003	0.002	0.003	−0.012	−0.039	−0.011	−0.039
	(0.024)	(0.024)	(0.01)	(0.01)	(0.027)	(0.025)	(0.029)	(0.026)
Product-market closeness ($t-1$)	0.841	0.852	0.25	0.262	0.708	0.88	0.801	0.939
	(0.443)	(0.445)	(0.303)	(0.306)	(0.449)	(0.491)	(0.479)	(0.511)
Exposure to the 2019 guidance	−0.000040	−0.000040	—	—	−0.000014	−0.000027	−0.000013	−0.000025
	(0.000018)	(0.000018)			(0.000017)	(0.000017)	(0.000018)	(0.000018)
Log first-time participants	0.01	0.012	0.002	0.002	0.002	0.004	0.002	0.004
	(0.008)	(0.008)	(0.006)	(0.006)	(0.008)	(0.007)	(0.008)	(0.007)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the firm-years where product-market closeness is defined, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B30: The firm's code and the industry's patents, $D(O_{it}, P_{G,t})$, on the short list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.02	0.02	0.012	0.012	0.014	0.016	0.015	0.016
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.003)	(0.003)
Log events per repository	−0.004	−0.005	−0.003	−0.003	−0.007	−0.006	−0.007	−0.006
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Log patents	0.001	0.002	0.001	0.001	0.002	0.003	0.002	0.003
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
Log assets	−0.003	−0.003	0.00044	−0.000060	0.001	0.00023	0.001	0.00015
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
R&D intensity	0.024	0.017	0.001	0.002	0.028	0.039	0.028	0.038
	(0.031)	(0.031)	(0.014)	(0.013)	(0.033)	(0.033)	(0.035)	(0.035)
Panel age	−0.002	−0.003	0.001	—	−0.003	−0.003	−0.003	−0.003
	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
B. Observed								
Log repositories	−0.047	−0.047	−0.016	−0.016	−0.029	−0.031	−0.03	−0.031
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.003)	(0.003)
Log events per repository	−0.004	−0.004	0.005	0.005	0.004	0.005	0.004	0.005
	(0.004)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Log patents	0.007	0.008	0.002	0.002	−0.000028	−0.003	−0.000055	−0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log assets	0.01	0.01	0.002	0.002	0.002	0.003	0.002	0.003
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
R&D intensity	0.091	0.082	−0.001	0.003	0.039	0.031	0.038	0.031
	(0.043)	(0.044)	(0.012)	(0.012)	(0.032)	(0.032)	(0.034)	(0.034)
Panel age	0.00050	−0.001	0.00026	—	−0.002	−0.001	−0.002	−0.001
	(0.001)	(0.002)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B31: The firm's code and the industry's patents, $D(O_{it}, P_{G,t})$, on the full list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.02	0.02	0.013	0.013	0.016	0.017	0.016	0.017
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log events per repository	−0.005	−0.006	−0.001	−0.002	−0.007	−0.006	−0.007	−0.006
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Log patents	−0.002	−0.001	0.003	0.003	0.003	0.002	0.003	0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log assets	−0.001	−0.002	0.003	0.003	0.002	0.002	0.002	0.001
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
R&D intensity	0.027	0.017	0.001	0.001	0.035	0.04	0.036	0.04
	(0.036)	(0.036)	(0.012)	(0.012)	(0.036)	(0.036)	(0.039)	(0.038)
Panel age	−0.002	−0.003	0.001	—	−0.003	−0.003	−0.003	−0.003
	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
Leader (top third, R&D rank)	0.001	0.002	−0.007	−0.006	−0.01	−0.012	−0.011	−0.012
	(0.012)	(0.012)	(0.007)	(0.007)	(0.012)	(0.012)	(0.012)	(0.012)
Product-market closeness $(t-1)$	0.054	0.06	0.023	0.028	0.211	0.229	0.232	0.246
	(0.181)	(0.184)	(0.076)	(0.075)	(0.161)	(0.169)	(0.175)	(0.178)
Exposure to the 2019 guidance	0.0000033	0.0000033	—	—	−0.0000092	−0.0000043	−0.0000094	−0.0000039
	(0.0000090)	(0.0000089)			(0.000011)	(0.0000099)	(0.000012)	(0.000010)
Log first-time participants	−0.001	−0.000025	0.001	0.001	−0.001	−0.00048	−0.001	−0.00041
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
B. Observed								
Log repositories	−0.046	−0.047	−0.013	−0.014	−0.024	−0.024	−0.024	−0.024
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Log events per repository	0.001	0.00024	0.006	0.006	0.011	0.012	0.011	0.012
	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Log patents	0.015	0.016	0.006	0.006	0.00043	−0.005	0.001	−0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log assets	0.002	0.001	0.007	0.007	0.003	0.006	0.002	0.006
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
R&D intensity	0.048	0.041	0.004	0.008	0.027	0.05	0.021	0.048
	(0.044)	(0.044)	(0.01)	(0.01)	(0.033)	(0.032)	(0.037)	(0.034)
Panel age	0.001	0.001	−0.001	—	−0.001	−0.00044	−0.001	−0.00030
	(0.002)	(0.002)	(0.001)		(0.002)	(0.002)	(0.002)	(0.002)
Leader (top third, R&D rank)	−0.02	−0.019	−0.011	−0.011	−0.015	−0.02	−0.016	−0.02
	(0.013)	(0.013)	(0.008)	(0.008)	(0.012)	(0.013)	(0.013)	(0.013)
Product-market closeness $(t-1)$	0.427	0.43	−0.009	−0.002	0.101	0.003	0.14	0.02
	(0.249)	(0.248)	(0.086)	(0.085)	(0.179)	(0.197)	(0.195)	(0.209)
Exposure to the 2019 guidance	−0.000022	−0.000022	—	—	−0.0000050	−0.0000032	−0.0000046	−0.0000028
	(0.000013)	(0.000013)			(0.000011)	(0.000012)	(0.000012)	(0.000013)
Log first-time participants	−0.002	−0.001	−0.003	−0.003	−0.008	−0.007	−0.008	−0.007
	(0.005)	(0.005)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the firm-years where product-market closeness is defined, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B32: The firm's patents and the industry's code, $D(P_{it}, O_{G,t})$, on the short list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.002	0.002	0.005	0.005	−0.018	−0.014	−0.019	−0.014
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Log events per repository	0.006	0.007	−0.004	−0.004	−0.002	−0.002	−0.002	−0.002
	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Log patents	0.002	0.002	0.009	0.01	0.012	0.014	0.012	0.014
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Log assets	−0.02	−0.02	−0.005	−0.006	−0.013	−0.016	−0.013	−0.016
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
R&D intensity	−0.085	−0.083	−0.026	−0.026	−0.111	−0.062	−0.112	−0.064
	(0.077)	(0.078)	(0.019)	(0.018)	(0.069)	(0.069)	(0.073)	(0.072)
Panel age	0.001	0.002	0.001	—	0.001	0.001	0.001	0.002
	(0.002)	(0.003)	(0.001)		(0.002)	(0.002)	(0.002)	(0.002)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
B. Observed								
Log repositories	−0.042	−0.042	−0.002	−0.003	−0.02	−0.021	−0.02	−0.02
	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Log events per repository	−0.014	−0.014	−0.006	−0.006	−0.008	−0.006	−0.007	−0.006
	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)
Log patents	−0.004	−0.004	−0.014	−0.014	−0.009	−0.014	−0.01	−0.014
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Log assets	0.006	0.006	−0.01	−0.01	−0.011	−0.009	−0.011	−0.01
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
R&D intensity	−0.041	−0.04	−0.039	−0.037	−0.159	−0.14	−0.161	−0.142
	(0.06)	(0.063)	(0.023)	(0.023)	(0.072)	(0.072)	(0.076)	(0.075)
Panel age	0.001	0.001	−0.001	—	−0.001	0.00016	−0.001	0.00028
	(0.002)	(0.003)	(0.001)		(0.002)	(0.002)	(0.003)	(0.003)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B33: The firm's patents and the industry's code, $D(P_{it}, O_{G,t})$, on the full list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	0.008	0.007	0.006	0.006	−0.016	−0.014	−0.017	−0.014
	(0.008)	(0.008)	(0.005)	(0.005)	(0.007)	(0.008)	(0.008)	(0.008)
Log events per repository	0.014	0.013	−0.002	−0.002	0.004	0.005	0.003	0.006
	(0.007)	(0.007)	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
Log patents	0.005	0.006	0.011	0.011	0.028	0.029	0.029	0.03
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
Log assets	−0.018	−0.019	−0.001	−0.001	−0.014	−0.016	−0.014	−0.017
	(0.007)	(0.007)	(0.005)	(0.005)	(0.007)	(0.007)	(0.008)	(0.007)
R&D intensity	−0.132	−0.138	−0.025	−0.025	−0.155	−0.139	−0.16	−0.146
	(0.097)	(0.099)	(0.021)	(0.021)	(0.079)	(0.081)	(0.085)	(0.088)
Panel age	−0.001	−0.002	−0.001	—	−0.003	−0.002	−0.004	−0.002
	(0.003)	(0.004)	(0.002)		(0.003)	(0.003)	(0.003)	(0.003)
Leader (top third, R&D rank)	−0.012	−0.011	0.002	0.003	−0.049	−0.042	−0.052	−0.042
	(0.022)	(0.022)	(0.014)	(0.014)	(0.017)	(0.019)	(0.018)	(0.019)
Product-market closeness ($t-1$)	0.066	0.072	0.165	0.173	0.6	0.658	0.583	0.653
	(0.43)	(0.434)	(0.146)	(0.148)	(0.321)	(0.312)	(0.344)	(0.329)
Exposure to the 2019 guidance	−0.0000013	−0.0000013	—	—	−0.000034	−0.000032	−0.000036	−0.000032
	(0.000023)	(0.000023)			(0.000021)	(0.000021)	(0.000021)	(0.000021)
Log first-time participants	−0.009	−0.008	0.004	0.004	−0.005	−0.006	−0.005	−0.006
	(0.006)	(0.006)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
B. Observed								
Log repositories	−0.039	−0.039	−0.001	−0.002	−0.017	−0.017	−0.017	−0.017
	(0.008)	(0.008)	(0.005)	(0.005)	(0.007)	(0.008)	(0.008)	(0.008)
Log events per repository	−0.005	−0.005	−0.002	−0.002	0.001	0.002	0.00041	0.002
	(0.008)	(0.008)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.008)
Log patents	0.007	0.007	−0.016	−0.016	−0.002	−0.008	−0.001	−0.008
	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
Log assets	−0.001	−0.001	−0.006	−0.006	−0.009	−0.004	−0.009	−0.004
	(0.008)	(0.008)	(0.005)	(0.006)	(0.008)	(0.008)	(0.009)	(0.008)
R&D intensity	−0.113	−0.116	−0.035	−0.031	−0.197	−0.156	−0.206	−0.161
	(0.063)	(0.067)	(0.025)	(0.025)	(0.085)	(0.077)	(0.091)	(0.084)
Panel age	−0.00045	−0.001	−0.003	—	−0.004	−0.001	−0.004	−0.001
	(0.003)	(0.004)	(0.002)		(0.003)	(0.003)	(0.003)	(0.003)
Leader (top third, R&D rank)	−0.063	−0.063	−0.009	−0.009	−0.057	−0.065	−0.059	−0.065
	(0.017)	(0.017)	(0.012)	(0.012)	(0.017)	(0.017)	(0.018)	(0.018)
Product-market closeness ($t-1$)	0.78	0.781	0.047	0.044	0.417	0.21	0.429	0.209
	(0.338)	(0.339)	(0.164)	(0.165)	(0.308)	(0.316)	(0.327)	(0.333)
Exposure to the 2019 guidance	−0.000020	−0.000020	—	—	0.0000042	−0.0000023	0.0000024	−0.0000023
	(0.000022)	(0.000022)			(0.000025)	(0.000023)	(0.000026)	(0.000023)
Log first-time participants	−0.004	−0.003	0.004	0.005	−0.007	−0.007	−0.007	−0.008
	(0.007)	(0.007)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the firm-years where product-market closeness is defined, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B34: The firm's patents and the industry's patents, $D(P_{it}, P_{G,t})$, on the short list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	−0.011	−0.011	−0.002	−0.001	−0.013	−0.014	−0.013	−0.014
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Log events per repository	0.002	0.002	−0.004	−0.004	0.005	0.005	0.005	0.005
	(0.006)	(0.006)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Log patents	0.049	0.049	0.055	0.055	0.045	0.055	0.045	0.055
	(0.005)	(0.005)	(0.004)	(0.004)	(0.006)	(0.006)	(0.007)	(0.006)
Log assets	−0.02	−0.02	0.002	0.003	−0.015	−0.02	−0.015	−0.02
	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.006)	(0.008)	(0.006)
R&D intensity	0.039	0.037	0.011	0.008	0.04	0.017	0.041	0.021
	(0.057)	(0.058)	(0.018)	(0.017)	(0.06)	(0.057)	(0.064)	(0.059)
Panel age	0.003	0.002	0.00023	—	0.00024	0.000054	0.00024	0.000074
	(0.002)	(0.003)	(0.001)		(0.002)	(0.002)	(0.003)	(0.002)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
B. Observed								
Log repositories	−0.013	−0.013	−0.005	−0.005	−0.004	−0.007	−0.004	−0.007
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Log events per repository	−0.01	−0.01	−0.004	−0.004	−0.003	−0.001	−0.003	−0.001
	(0.006)	(0.006)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Log patents	−0.033	−0.032	−0.035	−0.035	−0.049	−0.041	−0.049	−0.041
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)
Log assets	−0.025	−0.025	−0.008	−0.007	−0.017	−0.021	−0.017	−0.021
	(0.006)	(0.006)	(0.005)	(0.005)	(0.007)	(0.005)	(0.007)	(0.006)
R&D intensity	−0.106	−0.111	−0.02	−0.017	−0.055	−0.067	−0.053	−0.063
	(0.057)	(0.059)	(0.018)	(0.018)	(0.054)	(0.051)	(0.057)	(0.053)
Panel age	0.004	0.003	0.002	—	0.00018	−0.000073	0.00017	−0.000054
	(0.002)	(0.003)	(0.001)		(0.002)	(0.002)	(0.002)	(0.002)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B35: The firm's patents and the industry's patents, $D(P_{it}, P_{G,t})$, on the full list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the industry-pool benchmark								
Log repositories	−0.016	−0.016	0.002	0.003	−0.014	−0.013	−0.014	−0.013
	(0.008)	(0.008)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)	(0.008)
Log events per repository	−0.007	−0.008	−0.002	−0.002	0.003	0.002	0.003	0.002
	(0.009)	(0.009)	(0.004)	(0.004)	(0.007)	(0.008)	(0.008)	(0.008)
Log patents	0.066	0.067	0.049	0.049	0.074	0.078	0.075	0.078
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Log assets	−0.035	−0.036	0.01	0.01	−0.032	−0.03	−0.032	−0.029
	(0.007)	(0.007)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)	(0.008)
R&D intensity	−0.064	−0.078	0.001	−0.001	−0.091	−0.087	−0.095	−0.084
	(0.058)	(0.06)	(0.014)	(0.014)	(0.052)	(0.057)	(0.056)	(0.06)
Panel age	0.001	0.00014	−0.000099	—	−0.007	−0.004	−0.007	−0.004
	(0.003)	(0.003)	(0.002)		(0.003)	(0.003)	(0.003)	(0.003)
Leader (top third, R&D rank)	0.027	0.028	0.018	0.018	0.018	0.016	0.014	0.013
	(0.02)	(0.021)	(0.011)	(0.011)	(0.018)	(0.019)	(0.019)	(0.019)
Product-market closeness $(t-1)$	0.4	0.407	0.099	0.095	0.372	0.572	0.394	0.588
	(0.404)	(0.404)	(0.149)	(0.151)	(0.476)	(0.479)	(0.513)	(0.501)
Exposure to the 2019 guidance	−0.000090	−0.000090	—	—	−0.000087	−0.00011	−0.000087	−0.00011
	(0.000023)	(0.000023)			(0.000024)	(0.000024)	(0.000025)	(0.000025)
Log first-time participants	0.01	0.011	0.003	0.002	0.003	0.001	0.003	0.001
	(0.007)	(0.007)	(0.003)	(0.003)	(0.006)	(0.006)	(0.007)	(0.007)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
B. Observed								
Log repositories	−0.027	−0.028	−0.003	−0.004	−0.01	−0.009	−0.01	−0.009
	(0.009)	(0.009)	(0.006)	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)
Log events per repository	−0.018	−0.019	−0.001	−0.001	−0.002	−0.001	−0.002	−0.00041
	(0.009)	(0.009)	(0.004)	(0.004)	(0.007)	(0.008)	(0.007)	(0.008)
Log patents	−0.022	−0.021	−0.045	−0.046	−0.031	−0.032	−0.031	−0.032
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Log assets	−0.04	−0.04	−0.002	−0.002	−0.025	−0.024	−0.026	−0.024
	(0.008)	(0.008)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)
R&D intensity	−0.184	−0.198	−0.019	−0.014	−0.13	−0.115	−0.135	−0.113
	(0.066)	(0.069)	(0.019)	(0.018)	(0.048)	(0.05)	(0.051)	(0.053)
Panel age	0.004	0.003	0.002	—	−0.005	−0.002	−0.005	−0.002
	(0.003)	(0.003)	(0.002)		(0.003)	(0.003)	(0.003)	(0.003)
Leader (top third, R&D rank)	−0.004	−0.003	0.011	0.011	−0.022	−0.019	−0.024	−0.021
	(0.022)	(0.022)	(0.012)	(0.012)	(0.015)	(0.017)	(0.016)	(0.018)
Product-market closeness $(t-1)$	0.276	0.282	0.035	0.035	0.303	0.357	0.347	0.389
	(0.41)	(0.41)	(0.138)	(0.135)	(0.425)	(0.444)	(0.458)	(0.466)
Exposure to the 2019 guidance	0.0000081	0.0000083	—	—	0.000017	0.000013	0.000017	0.000013
	(0.000018)	(0.000018)			(0.000016)	(0.000016)	(0.000016)	(0.000016)
Log first-time participants	0.013	0.014	0.005	0.005	0.002	−0.001	0.002	−0.001
	(0.007)	(0.007)	(0.002)	(0.003)	(0.005)	(0.006)	(0.006)	(0.006)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the industry-pool benchmark; panel B: the observed distance, on the firm-years where product-market closeness is defined, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B36: The industry's patents and its code, the firm left out, $D(P_{G,t}, O_{G,t})$, on the short list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the within-industry benchmark								
Log repositories	−0.004	−0.003	0.004	0.006	0.001	0.003	0.00023	0.003
	(0.003)	(0.003)	(0.005)	(0.005)	(0.002)	(0.003)	(0.002)	(0.003)
Log events per repository	−0.005	−0.004	0.004	0.004	0.003	−0.002	0.003	−0.001
	(0.004)	(0.004)	(0.003)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)
Log patents	−0.004	−0.006	−0.012	−0.012	−0.001	0.000028	−0.00011	0.00049
	(0.003)	(0.004)	(0.004)	(0.004)	(0.002)	(0.003)	(0.002)	(0.003)
Log assets	0.013	0.014	−0.001	0.001	−0.001	0.002	−0.001	0.002
	(0.005)	(0.005)	(0.004)	(0.004)	(0.002)	(0.003)	(0.002)	(0.003)
R&D intensity	0.192	0.219	0.031	0.02	0.01	0.027	0.008	0.029
	(0.079)	(0.08)	(0.023)	(0.016)	(0.013)	(0.021)	(0.013)	(0.021)
Panel age	−0.003	−0.00050	−0.015	—	−0.00010	0.000070	−0.00032	−0.000024
	(0.001)	(0.002)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
B. Observed								
Log repositories	−0.042	−0.042	−0.003	−0.003	0.00046	−0.006	0.00049	−0.006
	(0.005)	(0.005)	(0.002)	(0.002)	(0.001)	(0.003)	(0.001)	(0.003)
Log events per repository	−0.019	−0.019	−0.003	−0.003	−0.002	−0.002	−0.002	−0.002
	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.003)
Log patents	0.013	0.013	−0.001	−0.001	0.001	−0.005	0.001	−0.005
	(0.005)	(0.005)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.003)
Log assets	0.028	0.028	−0.002	−0.001	0.001	0.005	0.001	0.005
	(0.006)	(0.006)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)
R&D intensity	0.123	0.119	−0.001	−0.002	0.008	−0.017	0.009	−0.016
	(0.089)	(0.09)	(0.005)	(0.005)	(0.013)	(0.031)	(0.013)	(0.032)
Panel age	0.002	0.002	−0.001	—	0.00017	0.001	0.00017	0.002
	(0.002)	(0.003)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
Firm-years	1,356	1,356	1,356	1,356	1,356	1,356	1,356	1,356
Firms	336	336	336	336	336	336	336	336
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the within-industry benchmark; panel B: the observed distance, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B37: The industry's patents and its code, the firm left out, $D(P_{G,t}, O_{G,t})$, on the full list: every column (extends Table B25)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Excess over the within-industry benchmark								
Log repositories	−0.014	−0.012	−0.002	0.001	0.001	0.004	0.001	0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.001)	(0.004)	(0.001)	(0.004)
Log events per repository	−0.013	−0.01	0.00020	0.001	0.001	−0.002	0.001	−0.001
	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
Log patents	−0.002	−0.005	−0.014	−0.013	0.00033	0.001	0.001	0.001
	(0.005)	(0.005)	(0.005)	(0.005)	(0.002)	(0.004)	(0.002)	(0.004)
Log assets	0.012	0.014	0.000093	0.004	0.00026	0.006	−0.000021	0.006
	(0.006)	(0.006)	(0.006)	(0.005)	(0.002)	(0.004)	(0.002)	(0.004)
R&D intensity	0.132	0.166	0.023	0.016	0.01	0.022	0.008	0.025
	(0.067)	(0.07)	(0.02)	(0.016)	(0.013)	(0.019)	(0.011)	(0.019)
Panel age	−0.002	0.001	−0.015	—	−0.001	−0.000046	−0.001	−0.00021
	(0.002)	(0.002)	(0.002)		(0.00049)	(0.001)	(0.00039)	(0.001)
Leader (top third, R&D rank)	−0.001	−0.004	0.002	0.000032	−0.009	−0.02	−0.01	−0.019
	(0.01)	(0.01)	(0.006)	(0.005)	(0.006)	(0.008)	(0.005)	(0.008)
Product-market closeness $(t-1)$	1.029	1.007	0.238	0.184	0.086	0.391	0.012	0.342
	(0.297)	(0.312)	(0.122)	(0.13)	(0.071)	(0.14)	(0.053)	(0.135)
Exposure to the 2019 guidance	−0.000011	−0.000011	—	—	−0.0000060	−0.000023	−0.0000060	−0.000023
	(0.000011)	(0.000012)			(0.0000097)	(0.000010)	(0.000010)	(0.000011)
Log first-time participants	0.012	0.009	0.001	0.00050	0.001	0.001	0.001	0.00033
	(0.004)	(0.004)	(0.002)	(0.002)	(0.001)	(0.003)	(0.001)	(0.003)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
B. Observed								
Log repositories	−0.045	−0.045	−0.002	−0.002	0.00029	−0.003	0.00039	−0.003
	(0.008)	(0.008)	(0.002)	(0.002)	(0.001)	(0.005)	(0.001)	(0.005)
Log events per repository	−0.018	−0.018	−0.001	−0.001	−0.000082	−0.003	−0.00012	−0.003
	(0.008)	(0.008)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.004)
Log patents	0.027	0.026	−0.003	−0.003	0.001	−0.005	0.001	−0.005
	(0.007)	(0.007)	(0.001)	(0.001)	(0.001)	(0.005)	(0.001)	(0.005)
Log assets	0.016	0.016	−0.003	−0.00049	−0.004	0.005	−0.004	0.005
	(0.009)	(0.009)	(0.002)	(0.002)	(0.001)	(0.004)	(0.001)	(0.004)
R&D intensity	0.073	0.076	−0.005	−0.003	−0.024	0.003	−0.025	0.007
	(0.093)	(0.094)	(0.004)	(0.005)	(0.014)	(0.029)	(0.016)	(0.031)
Panel age	0.004	0.004	−0.001	—	0.00037	0.003	0.00029	0.003
	(0.003)	(0.004)	(0.001)		(0.00039)	(0.002)	(0.00038)	(0.002)
Leader (top third, R&D rank)	−0.046	−0.047	0.00024	−0.001	0.015	−0.002	0.015	−0.003
	(0.023)	(0.023)	(0.002)	(0.002)	(0.005)	(0.013)	(0.005)	(0.013)
Product-market closeness $(t-1)$	1.061	1.054	0.014	−0.017	0.028	−0.089	0.04	−0.095
	(0.488)	(0.492)	(0.044)	(0.04)	(0.057)	(0.195)	(0.059)	(0.205)
Exposure to the 2019 guidance	−0.000055	−0.000055	—	—	0.0000091	−0.0000083	0.0000087	−0.0000086
	(0.000021)	(0.000021)			(0.0000056)	(0.000012)	(0.0000058)	(0.000012)
Log first-time participants	0.007	0.007	0.00026	0.00043	−0.001	−0.00029	−0.001	−0.00025
	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.004)
Firm-years	851	851	851	851	851	851	851	851
Firms	218	218	218	218	218	218	218	218
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

Event weights; the firm's industry group in the year before, the firm left out. Panel A: the excess over the within-industry benchmark; panel B: the observed distance, on the firm-years where product-market closeness is defined, on the same firm-years and right-hand side. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb. Industry: NAICS four digits, three where thin.

Table B38: Open-source activity on the firm's position: every column (extends Table B6)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. The leader indicator								
Log patents	0.075	0.082	−0.01	−0.009	0.142	0.149	0.141	0.151
	(0.061)	(0.062)	(0.053)	(0.053)	(0.058)	(0.06)	(0.061)	(0.061)
Log assets	0.232	0.23	0.214	0.21	0.356	0.359	0.363	0.359
	(0.067)	(0.067)	(0.062)	(0.065)	(0.071)	(0.072)	(0.075)	(0.074)
R&D intensity	1.955	1.852	0.58	0.57	1.703	2.27	1.773	2.313
	(1.058)	(1.06)	(0.213)	(0.221)	(1.032)	(1.177)	(1.116)	(1.235)
Panel age	−0.062	−0.07	0.09	—	−0.042	−0.044	−0.042	−0.044
	(0.021)	(0.025)	(0.011)		(0.027)	(0.024)	(0.028)	(0.025)
Leader (top third, R&D rank)	0.317	0.319	−0.001	−0.004	0.077	0.057	0.069	0.055
	(0.17)	(0.171)	(0.047)	(0.048)	(0.181)	(0.172)	(0.19)	(0.176)
Firm-years	1,347	1,347	1,347	1,347	1,347	1,347	1,347	1,347
Firms	333	333	333	333	333	333	333	333
B. The R&D rank's percentile								
Log patents	0.069	0.076	−0.011	−0.01	0.136	0.147	0.135	0.15
	(0.061)	(0.062)	(0.052)	(0.053)	(0.059)	(0.06)	(0.062)	(0.062)
Log assets	0.2	0.197	0.209	0.206	0.311	0.343	0.317	0.344
	(0.073)	(0.073)	(0.072)	(0.075)	(0.076)	(0.077)	(0.08)	(0.08)
R&D intensity	1.74	1.624	0.565	0.557	1.411	2.179	1.472	2.224
	(1.057)	(1.058)	(0.228)	(0.236)	(1.002)	(1.192)	(1.086)	(1.254)
Panel age	−0.06	−0.068	0.09	—	−0.041	−0.044	−0.042	−0.044
	(0.021)	(0.025)	(0.012)		(0.027)	(0.025)	(0.028)	(0.025)
R&D rank's percentile	1.031	1.045	0.072	0.06	0.716	0.308	0.7	0.293
	(0.493)	(0.495)	(0.326)	(0.328)	(0.555)	(0.515)	(0.583)	(0.532)
Firm-years	1,347	1,347	1,347	1,347	1,347	1,347	1,347	1,347
Firms	333	333	333	333	333	333	333	333
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

The outcome is $\log(1 +$ the firm's open-source repositories in the five-year window ending $t)$. Controls in every column. The counts are taken over the five-year window ending $t - 1$, and size, in US dollars, and R&D intensity at $t - 1$. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb.

Table B39: Open-source activity on the industry's totals: every column (extends Table B6)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Repositories in the window ending t								
Log patents	0.215	0.212	0.061	0.064	0.153	0.166	0.153	0.165
	(0.057)	(0.058)	(0.068)	(0.068)	(0.063)	(0.063)	(0.065)	(0.065)
Log assets	0.276	0.278	0.167	0.164	0.352	0.374	0.354	0.376
	(0.063)	(0.063)	(0.085)	(0.089)	(0.073)	(0.074)	(0.077)	(0.076)
R&D intensity	1.083	1.131	0.358	0.339	1.358	1.58	1.391	1.595
	(0.994)	(1.025)	(0.121)	(0.121)	(1.093)	(1.187)	(1.153)	(1.227)
Panel age	−0.047	−0.044	0.04	—	−0.038	−0.045	−0.038	−0.046
	(0.024)	(0.027)	(0.017)		(0.03)	(0.027)	(0.032)	(0.028)
Log others' patent mass	−0.272	−0.273	−0.136	−0.138	−0.275	−0.149	−0.282	−0.15
	(0.061)	(0.062)	(0.077)	(0.077)	(0.224)	(0.142)	(0.249)	(0.147)
Log others' code mass	0.294	0.296	0.135	0.143	0.017	0.294	0.007	0.297
	(0.06)	(0.061)	(0.06)	(0.062)	(0.224)	(0.111)	(0.249)	(0.116)
Log others with code	0.131	0.128	0.19	0.152	0.528	0.019	0.612	0.013
	(0.155)	(0.155)	(0.163)	(0.169)	(0.721)	(0.205)	(0.844)	(0.211)
Firm-years	974	974	974	974	974	974	974	974
Firms	278	278	278	278	278	278	278	278
B. Repositories first active in $t+1$, fiscal years 2021–2023								
Log patents	0.142	0.134	−0.08	−0.069	0.082	0.071	0.082	0.071
	(0.065)	(0.065)	(0.134)	(0.134)	(0.074)	(0.073)	(0.077)	(0.075)
Log assets	0.375	0.38	0.065	0.031	0.436	0.504	0.44	0.507
	(0.072)	(0.072)	(0.125)	(0.126)	(0.08)	(0.081)	(0.083)	(0.083)
R&D intensity	3.256	3.508	−0.273	0.092	2.959	3.879	3.108	3.982
	(1.165)	(1.193)	(0.788)	(0.798)	(1.315)	(1.275)	(1.392)	(1.329)
Panel age	−0.055	−0.044	−0.093	—	−0.036	−0.054	−0.037	−0.055
	(0.029)	(0.031)	(0.052)		(0.036)	(0.032)	(0.037)	(0.033)
Log others' patent mass	−0.039	−0.044	−0.235	−0.304	−0.124	−0.058	−0.118	−0.05
	(0.069)	(0.069)	(0.26)	(0.266)	(0.219)	(0.156)	(0.241)	(0.16)
Log others' code mass	0.226	0.232	0.144	0.226	−0.08	0.312	−0.123	0.305
	(0.07)	(0.07)	(0.229)	(0.241)	(0.271)	(0.107)	(0.307)	(0.111)
Log others with code	0.14	0.133	0.351	0.172	0.484	−0.031	0.72	−0.031
	(0.181)	(0.181)	(0.691)	(0.696)	(1.018)	(0.214)	(1.181)	(0.22)
Firm-years	700	700	700	700	700	700	700	700
Firms	250	250	250	250	250	250	250	250
Firm			Yes	Yes				
Year		Yes		Yes	Yes	Yes		
Industry					Yes			
SIC2						Yes		
Industry × year							Yes	
SIC2 × year								Yes

The other members of the firm's industry group at $t-1$, the firm left out: their patent mass, their code mass under event weights and the number with code, each in logs. The counts are taken over the five-year window ending $t-1$, and size, in US dollars, and R&D intensity at $t-1$. Panel B's outcome ends in 2023 because the event data end in 2024. Standard errors clustered by firm, in parentheses; a dash marks a term the fixed effects absorb.

Table B40: The excess under other readings of the classification

Reading of the CPC	Firm-years	Observed distance	Benchmark distance	Excess	95% interval	Share outside the patent support	Ratio to the main-group excess [interval]
Main group, as recorded	2,472	0.8055	0.5719	0.2336	[0.2228, 0.2443]	0.857	—
Subclass	2,472	0.6326	0.3867	0.2459	[0.2318, 0.2599]	0.223	1.053 [1.011, 1.094]
Class	2,472	0.4924	0.2694	0.2231	[0.2062, 0.24]	0.129	0.955 [0.893, 1.018]
Without G06F 8/00	2,470	0.7891	0.5786	0.2105	[0.1997, 0.2212]	0.736	0.9 [0.884, 0.916]
G06F 8/00 to 11/00 merged	2,472	0.7978	0.5524	0.2455	[0.2345, 0.2564]	0.622	1.051 [1.043, 1.059]
One main group per item, drawn by weight	2,472	0.8383	0.6681	0.1702	[0.16, 0.1804]	1.241	0.729 [0.706, 0.751]
One main group per item, largest weight	2,472	0.8222	0.6152	0.207	[0.1949, 0.2191]	1.244	0.886 [0.858, 0.915]

CPC is the Cooperative Patent Classification. Event weights; the comparison sample. Each row reads both channels at its own level of the classification and uses the same draws of the randomization benchmark. The benchmark distance is the distance's mean over the benchmark's draws. The excess is the observed distance minus the benchmark distance, with a 95 percent interval clustered by firm. The firm's patent support is the technologies, at the row's level, where it holds patents in the same window. The share outside the patent support is the part of the excess in cells outside the firm's patent support. A cell is one firm in one technology, at the row's level of the classification, in one year. A share above one means that the excess inside the patent support is negative: there the firm's open-source code sits nearer its patents than the randomization benchmark places it. The ratio compares a row's excess with the main-group excess on the same firm-years. It is a description, since the supports differ. The rows with one main group per item give every patent and every open-source project a single main group, so that both channels have the same breadth. The row drawn by weight draws that main group once, with probabilities equal to the item's weights over its main groups. Two more such draws give excesses of 0.1708 and 0.1673, both inside the first draw's interval. The last row takes the main group with the largest weight. Equal weights give the same pattern.

that the randomization benchmark alone produces, and subtract the draws' mean from the observed coefficient. The difference is the excess's coefficient: what goes with the excess. The spread of the draws' coefficients shows how far the randomization benchmark alone could move the coefficient.

Each comparison reads its observed coefficient against that coefficient's mean over the benchmark's draws. With the index k running over the four comparisons, the regression and the difference for comparison k are

$$\begin{aligned} D_{it}^{(d)} &= \theta_k^{(d)}\, z_{i,t-1}^k + \beta_R \log(1 + n_{i,t-1}^{\text{oss}}) + \beta_E \log\Big(1 + \frac{n_{i,t-1}^{\text{ev}}}{n_{i,t-1}^{\text{oss}}}\Big) + \beta_P \log(1 + n_{i,t-1}^{\text{pat}}) \\ &\quad + \phi_{it} + \varepsilon_{it}, \qquad d \in \{\text{obs}, 1, \ldots, K\}, \\ \Delta_k &= \theta_k^{\text{obs}} - \frac{1}{K}\sum_{d=1}^{K} \theta_k^{(d)}. \end{aligned} \tag{6}$$

The index d names the outcome. $D_{it}^{(\text{obs})}$ is the observed distance D_{it} between firm i's open-source code and its patents over the five years ending in year t, and $D_{it}^{(1)}, \ldots, D_{it}^{(K)}$ are the distances in the benchmark's draws, whose mean is the benchmark distance μ_{it}. The randomization benchmark has K draws, and K is 200, a number fixed before the run. $z_{i,t-1}^k$ is the comparison's variable in the year before. For the count slope it is the log of one plus the count of open-source projects, and this term then appears once; for the other three it is R&D intensity, log total assets on the US-dollar firm-years, or indicators for six sectors, the two-digit NAICS codes, against NAICS 51. $\theta_k^{(d)}$ is the coefficient on $z_{i,t-1}^k$ in the regression on outcome d, so θ_k^{obs} is the observed coefficient and

$\theta_k^{(1)}, \ldots, \theta_k^{(K)}$ are the draws'. The counts are those of equation (1): $n_{i,t-1}^{\text{oss}}$, $n_{i,t-1}^{\text{ev}}$ and $n_{i,t-1}^{\text{pat}}$ are the firm's open-source projects, open-source events and patents over the five years that end the year before, and β_R, β_E and β_P are the slopes on the three log terms. For the count slope, ϕ_{it} is firm and year effects, so the comparison sets each firm beside itself over time; the other three comparisons have no fixed effects and set firms beside one another. ε_{it} is the error. Δ_k is the observed coefficient minus the draws' mean. Since the benchmark distance is the draws' mean and least squares is linear in the outcome, Δ_k equals the coefficient on $z_{i,t-1}^k$ with the excess on the left, the excess's coefficient.

The method is randomization inference: it re-estimates the same coefficient on each draw of the randomization benchmark and reads the observed one against them (Young, 2019, pp. 572–573). Duranton and Overman (2005, LSE Research Online version, p. 10) recompute their statistic on each simulated location pattern, with as many sites as the industry has plants, and Gentzkow et al. (2019b, p. 1318) re-estimate their measure on a permutation of the group labels. Subtracting the benchmark distance follows Carrington and Troske (1997, pp. 405–406), who measure segregation by its excess over the level that random allocation produces.

Before testing, I declared that a comparison goes with the excess when two things hold: the observed coefficient lies outside the central 95 percent of the draws' coefficients, and two p-values stay at or below 0.05 after Holm's adjustment (Holm, 1979). The first p-value is the draws': it counts the draws with a coefficient at least as far from the draws' mean as the observed one, adds one to that count and divides by $K + 1$. The second is the p-value of Δ_k, from standard errors clustered by firm. For industry, I read the six coefficients together, by how far they lie jointly from the draws' mean. Four comparisons give four tries, each of which could show a difference at random, and Holm's adjustment accounts for that: it multiplies the smallest of the four p-values by four, the next by three, and so on. If no comparison goes with the excess, the adjustment keeps the probability of reading any of them as going with it at 5 percent or less. With K draws, the smallest p-value the draws can give is $1/(K + 1)$, which at 200 draws stays below 0.05 even when multiplied by four.

Of the three comparisons beside the count slope (section 5.3), only industry goes with the excess, jointly across the six sectors. Size and R&D intensity show no clear link with the excess: under event weights, their Δ_k is 0.0018 (0.0033) for log total assets and 0.0688 (0.0601) for R&D intensity, and neither survives the adjustment. For R&D intensity, the excess's coefficient is larger than for the count slope, 0.0688 against 0.0516, but far less precise. In economic terms, firms in different industries differ in how much their open-source code and patents are separated beyond the randomization benchmark, while larger firms and more research-intensive firms show no clear difference.

Next to the seniority comparisons of section 5.3, I read the firms' patents against two stated predictions of Sørensen and Stuart (2000, p. 88), that older firms build more on their own earlier innovations and refine older areas of technology. They measure the first by the rate of self-citing patents (p. 92) and the second by the lag of the cited prior art (p. 93). Their prediction concerns

Table B41: The four comparisons declared before testing, against the randomization benchmark

Weights	Comparison	Firm-years	Observed	Benchmark	Difference (SE)	Holm-adjusted p-values: draws; difference	After Holm
Equal	Count slope	2,472	-0.0204^{***}	-0.0832	0.0628^{***} (0.0069)	0.02; < 0.001	beyond
	R&D intensity	1,998	-0.285^{***}	-0.3142	0.0293 (0.0617)	0.418; 1.0	not distinguishable
	Log total assets	1,743	0.016^{***}	0.0152	0.0008 (0.0033)	0.592; 1.0	not distinguishable
	Industry, joint	2,409	M 229.7	q_{95} 11.7	F 7.79	0.02; < 0.001	beyond
Event	Count slope	2,472	-0.0152^{***}	-0.0668	0.0516^{***} (0.0072)	0.02; < 0.001	beyond
	R&D intensity	1,998	-0.2263^{***}	-0.2951	0.0688 (0.0601)	0.07; 0.505	not distinguishable
	Log total assets	1,743	0.0171^{***}	0.0153	0.0018 (0.0033)	0.318; 0.579	not distinguishable
	Industry, joint	2,409	M 98.2	q_{95} 11.9	F 6.67	0.02; < 0.001	beyond

OSS means open-source software. Event weights count each OSS project by its events. Equal weights count each project once. Each comparison uses the comparison sample's firm-years that have its variable. It runs the same regression on the observed distance and on each of the benchmark's draws. The right-hand side holds the log counts of OSS projects, events per OSS project and patents over the five-year window that ends the year before. The count slope's variable is the first of these, the log of one plus the count of OSS projects. Observed is the coefficient on the comparison's variable in the regression on the observed distance. Benchmark is the same coefficient's mean over the benchmark's draws. Difference is Observed minus Benchmark, the excess's coefficient, with its standard error. In the industry rows, these three columns hold the joint statistics defined below. The count slope has firm and year effects, and the other three comparisons have none. R&D intensity and log total assets come from the year before. The comparison on log total assets uses the US-dollar firm-years only. Total assets are in each firm's reporting currency, so their levels are comparable only among firms that report in US dollars. Industry is six sectors, the two-digit NAICS codes, against NAICS 51, read jointly. For industry, M is the Mahalanobis distance of the observed coefficients from the draws' mean. q_{95} is the 95th percentile of the same distance among the draws. F is the clustered F-statistic of the difference. Holm's adjustment (Holm, 1979) runs across the four comparisons, for the draws' p-value and for the clustered p-value of the difference. "Beyond" means that the observed coefficient lies outside the draws' central 95 percent and both p-values survive the adjustment. "Not distinguishable" means that they do not. Standard errors are clustered by firm. Stars mark the observed coefficient and the difference, from their clustered p-values before Holm's adjustment: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from standard errors clustered by firm. No star marks the randomization benchmark's column, the Holm-adjusted p-values or industry's joint statistics.

the relative tendency to self-cite, with no strong prediction for other patents (p. 92). On the 381 firms whose age is measured, mostly listed in the United States, the share of patents that cite the firm's own patents and the mean citation lag both have the predicted positive sign on age. Neither coefficient is distinguishable from zero ($p = 0.673$ and 0.106).

B.12 The count slope against a mix of the firm's own

This appendix reads the count slope of section 5.2, the slope on the log count of a firm's open-source projects, against simulated firms that keep one technology mix of their own at every size, and finds it close to theirs. Section 5.2 reads the observed distance's slope against 1,000 further draws of the benchmark (Figure 3); this appendix reads the excess's slope.

A firm whose open-source code keeps one technology mix of its own at every size produces an excess whose count slope is close to the observed one. In the simulated firms, each firm's patents are as observed, and its open-source projects receive the technologies of the firm's own projects in a random order, so that each firm-year's open-source code is drawn, at its observed count, from the mix of all the firm's projects. The counts, the years each project is active and the event weights stay as observed. Unlike the simulated firms of section 4.2, in which technology plays no part in what a firm opens, these keep the difference between the mix of each firm's open-source code and the mix of its patents. For each of 200 sets of simulated firms, I draw a randomization benchmark of its own and re-estimate columns 4 and 6 of Table 7. Since the excess is the observed distance minus the benchmark distance, its slope is the observed slope minus the benchmark distance's slope. In a firm with twice the open-source projects of another firm in its industry-year, the excess is higher by 0.0234 in the data and by 0.0221 on average in the simulated firms ($p = 0.299$); when a firm's own projects double, it rises by 0.0357 in the data and by 0.0324 in the simulated firms ($p = 0.169$). Each p is the two-sided Monte Carlo p-value of the data's slope among the sets' slopes, so a large one means that the simulated firms reproduce the data (Table B42).

In such firms, the slope comes from the level of the excess alone, since the mix of a firm's open-source code does not change as its program grows. A larger program takes more projects from the same mix, so its distance from the firm's patents falls only as the small-unit bias fades (section 5.2). Open-source code whose mix differs from the patents' responds less to that bias than open-source code whose technologies are dealt out again at random. The excess therefore rises with the count whenever a firm's open-source code and patents keep different mixes, and section 4.3 shows that they lie further apart than the randomization benchmark places them in nearly every firm.

The simulated firms also break the link between a project's technologies and its start year, its years of activity and its event weight. Agreement with them shows that the data are consistent with each firm keeping its mix at every size; it does not show that each firm does.

Table B42: The count slope in the data and in simulated firms that keep a mix of their own

	Data	Simulated firms, mean	Simulated firms, central 95%	p
Across firms in an industry-year				
Distance	−0.0261	−0.0289	−0.0312 to −0.0259	0.070
Benchmark distance	−0.0495	−0.0510	−0.0517 to −0.0503	0.010
Excess	0.0234	0.0221	0.0196 to 0.0248	0.299
Within a firm				
Distance	−0.0104	−0.0146	−0.0205 to −0.0103	0.070
Benchmark distance	−0.0461	−0.0470	−0.0486 to −0.0455	0.219
Excess	0.0357	0.0324	0.0265 to 0.0370	0.169

Each entry is the coefficient on log one plus the firm's lagged open-source projects, times $\ln 2$: the change when a firm's open-source projects double. The distance is D_{it} under event weights; the benchmark distance is its mean over the randomization benchmark; the excess is the first minus the second. Across firms: industry-by-year effects, with the share of the firm's patents outside the classes that carry open-source code; within a firm: firm and year effects. The simulated firms' columns summarize 200 sets of simulated firms. p is the two-sided Monte Carlo p-value of the data's slope among the sets' slopes. The distance and the excess are the tests, read with Holm's adjustment across the two within each panel; the benchmark distance's row describes.

C The data and the validation of the labels

This appendix gives the validation of the repository labels of section 3.3 and the run that measured their reproducibility, then what the labeler reads of a long README and the limits of the patent crosswalk.

Validation is measured against references that are themselves automated, and the paper says so. Label agreement is scored against the EPO's text categorizer, which serves as a check rather than as the labeler because of its limits on input length and time (section 1): at the main group, Cohen's κ (Cohen, 1960, p. 40), agreement net of chance, is 0.3444 (standard error 0.0248) on READMEs, the text the labeler reads, on 432 items, where the two primary codes agree on 40.05%, and 0.4521 on abstract-derived text, a patent-style abstract that a language model writes from the repository's full README so that the categorizer, which reads patent prose, can classify it. Automated CPC classification has been benchmarked on patents, from a fine-tuned BERT model at subclass (Lee and Hsiang, 2020, pp. 1–3) to encoders and open-weight language models whose evaluation stops at subclass (Emer et al., 2026, pp. 1–3, 21); the labels here classify repositories, at main group. No human-coded CPC gold exists, and agreement is not accuracy: two automated systems can agree for shared reasons, prompted language models label below the best fine-tuned classifiers while reaching fair agreement with human labels (Ziems et al., 2024, pp. 238, 240), and a measurement built with a language model needs a validation sample drawn from the data it measures, as Dell and Rambachan (2026) argue. Such a sample serves to correct the downstream estimates, not only to report agreement on it (Duan and Pelger, 2026, §1), since validation data can remove the bias that measurement error brings into estimates built on predicted labels (Ludwig et al., 2026, §4.3). A pilot of a human-coded probability sample has been coded and scored: on 50 items the expert's coding shares at least one main group with production in 34, carries 0.49 of production's label mass, and equals the production set in 9 items. That is agreement with one expert's coding on one sample, not an accuracy rate and not a gold standard: there is a single coder, the coder is the author, and no two-coder reliability is available. The EPO comparison has been widened to 800 repositories, the pilot's items and new ones drawn by the pilot's rule, classified through the categorizer's web page. A scoring rule declared before any score was read treats an item for which the categorizer returns no code as empty: it is counted and left out of the pairwise measures, never scored as disagreement. The categorizer returned a code for 91% of the items. On the 721 items where both return codes, the production labels share a main group with the categorizer on 76.4% of items (95 percent Wilson interval 73.2 to 79.4 percent) and a subclass on 90.4%; they agree on the exact set on 10.7%, with a mean Jaccard overlap of 0.369, and 56.4% of the production labels' mass falls inside the categorizer's set. The disagreement lies mostly within a subclass. The model lists its main groups in order, and its first is the categorizer's first code on 43.9% of the 716 items whose ordered answer matches the production labels, against 14.5% when the production label sets are dealt out at random among the items in the same CPC section as the categorizer's first code. The second model, a distinct anonymous model that is no longer listed, labels agree less (a shared main group on 43.3%, a mean Jaccard overlap of 0.208), and the author's own coding of the pilot

items sits between the two (a shared main group on 63.4%). This is agreement with an automated reference, not an accuracy rate and not a gold standard. The paper's placement statements rest on the assumption stated here: that label errors are unrelated to the firm's patent technologies and to the channel. The identity-leakage check, an invariance test that changes a label-irrelevant input such as a name and expects the prediction to stay (Ribeiro et al., 2020, §2.1), then asked whether the labeler places a repository in its owner's patent technologies for reasons unrelated to what the code does: on a frozen probability sample of 7,600 repositories, masking the owner's name moves the in-support share of label mass by 0.0011 (95 percent interval −0.002 to 0.0042), against a predeclared materiality rule that the run-to-run arm's own upper bound of 0.0041 also clears. The verdict reads *clean at this resolution.* Masking owner names inside the repository text as well moves the share by 0.0037 — the one contrast whose interval excludes zero, still an order of magnitude below the rule, reported as a small association carried by label churn rather than material pull. Two limits are declared with the verdict: it is not proof of no training exposure, which text posted after the language model's training would address (Gilardi et al., 2023; Ludwig et al., 2026), and it ran under the production model's second served name. The contrast that would address training exposure, agreement split by whether a repository was created before or after the language model's training cutoff, has not been run: the provider does not state the language model's training cutoff, so any split would rest on an assumed date. It would run only on the author's word. A set of perturbations declared in advance then closed the third: five prompt and menu variants, invariance tests of the same kind (Ribeiro et al., 2020, §2.1) and the sensitivity analysis to prompt design that Carlson and Burbano (2026, p. 699) propose, moved the in-support share by no more than 0.0071 on the battery's model and 0.0076 on a second, anonymous model, with every interval covering zero — the labels do not depend on the prompt's wording or the menu's order, though dropping the menu's titles churns 63.5% of label sets. The same battery's anchor quantifies what a model is: on identical prompts the anonymous model sits 0.0577 below the production labels in in-support share and reproduces 16.4% of their label sets, against the same model's re-run's three-fifths. This battery measures formulation dependence; it does not measure accuracy.

The same run measured reproducibility, and its finding shapes how the labels may be used. Re-running the identical prompts reproduces 60% of production label sets exactly, while about 40% of label sets change under every perturbation — including a pure re-run, and including a mask that removes no owner identity at all. Label-level objects churn; portfolio-level objects do not: a repository's in-support share is exactly unchanged in at least nine pairs in ten under every perturbation, because owners' patent technologies are broad and the churn stays largely inside them, so the shares, the in-support measure and the level of the distance are the stable objects. A re-labeling is therefore a new set of labels rather than a reproduction, and this paper leans on the stable, portfolio-level objects. The labels also remain one retrospective classification pass: they read current repository text and apply it to the whole event history. What is dated is the activity, its events, the entry of each repository and the order of the two channels; what a repository's code is about is not, though code changes over a repository's life (section 6).

A second set of labels then tests the facts rather than the labels themselves. A complete second set of labels from another model, given the same prompt and option menu, replaces production's repository labels; the panel, the windows, the patent side, the events and the distance stay the record's. Repositories that only production labels carry 29.5% of the firm-year panel's open-source events and leave; those that only the second model labels carry 0.04%. Under event weights the mean distance is 0.8548 with the second model's labels against 0.8089 with production's, a difference of 0.046 (firm-clustered standard error 0.0035). Across firm-years the two distances correlate at 0.814 under event weights and 0.909 under equal weights. The second model places less of its label mass where the owner patents, which is the direction of the change. Keeping production's labels where the second model leaves a repository empty gives a difference of 0.0315. The excess over the randomization benchmark, on the comparison sample's firm-years with the seed, draws and firm order of the headline excess, is 0.3091 under equal weights and 0.2205 under event weights, against 0.3304 and 0.2337 for the record; the excess is a finding under both weights, with the second model's labels alone and with production's where it is empty, each within the declared distance of the record's. Two models agreeing on the facts is robustness to which model labels the repositories, not evidence that either labels them correctly. Both read the same prompt over the same menu, so what they share is not tested, and production's labels stay the labels the paper uses. Under the second model's labels alone, 180 firm-years of the panel keep fewer than five labeled repositories; they stay in, as the panel's floor of five repositories is not re-applied.

Two further checks read the excess where the labels may be weakest. At the repository level, the share of a repository's code outside the firm's patent support exceeds its value under the randomization benchmark by 0.0711 on all repositories. Among the repositories in the re-run sample, it exceeds it by 0.0868 on those whose labels a second run of the same model reproduces. On repositories outside the quarter with the shortest README text, it exceeds it by 0.0801. A different model, which chooses main groups within the same subclasses as the production model, keeps 96% of the excess under event weights and 96% under equal weights. Where the excess lies depends more on the labels than its size does. With that model's labels and each firm-year counted once, the share of the excess outside the patent support is 97.3% under event weights, against 85.7% with the production labels, and 70.5% against 63% under equal weights.

What the labeler reads. The pipeline caps a repository's cleaned README at 12,000 characters, and the model reads its first 6,000. The model's cap truncates the README of 20.4% of labeled repositories, and the median of those keeps 60.9% of its text; 7.3% reach the pipeline's cap. Whether the cap moves the portfolio-level objects is untested, and classifying long READMEs in full, by word frequencies or embeddings, would show it (section 6).

The patent crosswalk. On the current crosswalk's full ledger of links, with no restriction on grant year, 22,743 patents are attached to two or three firms of the comparison sample, mostly corporate-family duplicates that the current convention counts in full for each record; on the bridge the distances read, grants of 2015–2024, 12,844 are (Appendix B.3). The panel dates ownership, as

a dated crosswalk does (Arora et al., 2021; Kogan et al., 2017). Each patent belongs to the firm that owns it in its grant year, and each repository-year to the firm that owns the repository in that year. A firm's record ends with its last year in Compustat, whatever the reason, and the panel holds none of its later firm-years. When the record ended in an acquisition or a merger, its later patents and repository-years count for the acquirer that a dated record of a majority acquisition names within a year of the record's end, and otherwise for no firm. A move dated only by the record's end in Compustat is made only when a second source names the same acquirer, and an acquirer whose own record had ended is followed to its owner only where that ownership is shown.

Patents granted in 2021–2024 enter only the windows that end after 2020. Across firm-years, with year effects, the observed distance falls with the share of the window's patents granted in 2021–2024 (coefficient -0.175, standard error 0.0405), while the excess under event weights shows no clear association with that share (coefficient -0.0376, $p = 0.262$). On the firm-years whose windows end in 2020, before these grants enter, the comparison with the ten nearest technology neighbors (Appendix E.8) keeps its Holm-adjusted p of 0.0199, against 0.0199 on all firm-years.

C.1 Methods for labeling open-source code

Table C1 compares the methods that could place an open-source project in the patent classification.

Table C1: Methods for assigning patent classes to an open-source project

Method	Inputs and preparation	Coverage of varied projects	Cost at this scale
Keyword rules	A list of words for each of the 9,867 main groups	Misses projects that describe the same technology in other words	Experts writing and testing every list
Hand coding	Coders who know the patent classification	Reads any project	An expert reading each of the 209,351 projects
A classifier trained on hand labels	Projects already labeled in the patent classification, which would have to be built first	Covers only what its training set covers	The hand labels, before any training
Transfer from patents (Ganglmair and Kann, 2025b)	Patent text to learn from	Needs a text that describes one invention, which a project seldom has, and stops at the subclass	Low once trained
Software categories (Brea, 2024; Conti et al., 2025)	A list of software categories	Coarse software categories, with no patent class, and Conti et al. (2025) set aside 18 percent of projects as uncategorizable	Low
The EPO's text categorizer	Nothing from the user. It learns from examiners' classifications	Built for patent prose, and reads at most 5,000 characters of a text	About 21 days for this data, behind captchas
A prompted language model (this paper)	The titles of the patent classification, given in the prompt	Reads any project's name, description, topics and README	About an hour, at a low cost at list prices

Notes: Each row describes a method for assigning classes from the Cooperative Patent Classification to an open-source project. The EPO's time reflects throughput in my runs. The language model's cost is an estimate at list prices for the corpus of September 2026, not a billed amount. This paper uses the classification at the level of the main group (section 3.3).

C.2 The second label check: another model on patents

Another language model, DeepSeek-V4-Flash, applies the same two stages to the titles and abstracts of 2,802 patents and labels 2,774 of them. The examiners' classifications are the reference. Table C2 compares the two at the main group and at the subclass, each against the model's label sets permuted across the patents within the CPC section of the examiners' primary code. Each patent is weighted by the inverse of its probability of entering the draw. This is agreement with the examiners

on patents, not accuracy, and it does not say how production's model reads code.

Table C2: The second label check: another model's labels for patents against the examiners'

	Main group		Subclass	
Agreement	Observed	At random [band]	Observed	At random [band]
A common group	89.1%	20.8% [18%, 23.6%]	95.4%	51.3% [48.7%, 54%]
The same first code	47.8%	5.5% [4%, 7.3%]	70.4%	29.1% [26.1%, 31.6%]
The same set	9.6%	0.5% [0.1%, 1%]	39.8%	10.9% [8.6%, 13%]
Mean Jaccard overlap	0.39	0.06 [0.05, 0.07]	0.64	0.27 [0.25, 0.29]
Cohen's κ, first codes	0.46	0.03 [0.01, 0.05]	0.63	0.15 [0.12, 0.18]

Notes: The labels are DeepSeek-V4-Flash's for 2,774 patents, and the reference is the examiners' classification of the same patents. At random, the model's label sets are permuted across the patents within the CPC section of the examiners' primary code. Each At-random column gives the draws' mean, with their 2.5th and 97.5th percentiles in brackets. Cohen's κ compares each patent's first code. The model lists at most three main groups, which bounds agreement on the same set.

C.3 The sample, the labeler and the labels' limits

This subsection gives the detail behind section 3: the sample against Compustat North America and the patent bridge, the public code's licenses and what falls outside the measure, the labeler's settings, and the labels' limits. Table C3 gives the counts and the firm's own distance in the comparison sample.

Table C3: Counts and the firm's own distance in the comparison sample

	Mean	SD	10th	25th	Median	75th	90th
Panel A. Counts over the firm-year's five-year window							
Open-source projects	200.4	820.2	8.0	14.0	41.0	125.0	345.9
Events on open-source projects	165,616	1,107,273	331	1,341	6,908	37,216	166,672
Events per open-source project	408.0	1555.1	26.4	58.3	159.0	401.7	920.0
Patents	1012.2	3377.4	10.0	21.0	83.5	520.2	2079.8
Main groups with the firm's open-source code	26.4	29.3	7.0	11.0	18.0	31.0	49.9
Main groups with patents	171.1	312.1	12.0	21.0	49.0	158.0	463.0
Main groups with both	15.4	22.1	2.0	4.0	9.0	17.0	34.0
Panel B. The firm's own distance, event weights							
Distance	0.8055	0.1649	0.5595	0.692	0.844	0.9504	0.9924
The benchmark distance	0.5719	0.1774	0.3368	0.4373	0.562	0.7142	0.8149
The excess	0.2336	0.1356	0.061	0.1379	0.2358	0.33	0.4097

Notes: SD is the standard deviation. The comparison sample has 2,472 firm-years from 595 firms, with each firm-year covering its five-year window. The open-source projects have a main group, belong to accounts linked to the 2024 parent, are owned by the firm in the year of their events, and have events in the window. The events are GitHub Archive events of every type and from every account, with corrected event dates. The patents are grants in the window that the patent–firm crosswalk links to the firm. A main group counts for a channel when the channel has some weight there. The observed distance is the Bar–Leiponen distance between the firm's open-source code and its patents. The benchmark distance is the mean distance over the benchmark's draws. The excess is the observed distance minus the benchmark distance. Across the sample's firm-years, the firms hold 137,510 distinct open-source projects and 885,230 distinct patents. In 4.98% of firm-years, the two channels share no main group. The columns from 10th to 90th give percentiles.

Table C4 follows the firm-years from the comparison sample to each regression's sample.

Table C4: The sample flow

			Excluded		
Step	Firm-years	Firms	Firm-years	Firms	Exclusion
Panel A. Each step keeps the previous step's firm-years with the named variable					
The comparison sample	2,472	595			
With an industry	2,450	594	22	1	Firm-years with no NAICS code in the year before
Panel B. The regression samples					
Section 5.2	2,447	589	25		From the comparison sample, the firm-years alone in a fixed-effect group
Section 5.5	2,413	585	26		From the 2,439 firm-years with an industry and all six distances, those alone in a fixed-effect group
Appendix B.9	883	226	84		From the 967 firm-years with every control of Appendix B.9, those alone in a fixed-effect group

Notes: NAICS is the North American Industry Classification System. Panel A begins with the comparison sample and requires each variable in turn. Panel B reports the regression samples. I exclude a firm-year that is alone in a fixed-effect group because its fixed effect fits it exactly. The main regressions in section 5 retain every firm-year identified by their fixed effects. They use 2,447 firm-years of 589 firms. The regressions in section 5.5 retain every firm-year with an industry and all six distances that their fixed effects leave identified. They use 2,413 firm-years of 585 firms.

The sample. The sample is large firms: the parents of Compustat firms to which patents and repositories link, a parent being the company at the top of the ownership tree, to which its subsidiaries' patents and repositories are assigned. In the panel's years the median parent firm-year is larger in total assets than 79.4% of the same year's Compustat North America firms; 79.1% of panel firm-years report research and development, against 25.9% of the universe; and business services alone hold 44.7% of the panel firm-years that carry a sector code. The findings speak to large, listed firms, 76.3% of them incorporated in the United States and the rest foreign firms carried in Compustat North America, and only to the organizations that link to a parent; firms outside Compustat, code published by individuals, and code that a firm's own employees release under personal accounts are outside it, much as anonymous research disclosures carry no firm (Ganglmair and Kann, 2025a, slide 6). Of the 4,333 listed patenting firms at risk of beginning to publish code, not yet publishing when observation begins, 317 (7.3%) begin to publish linked code in 2016–2024. The sample is also the chance the paper takes: these are the cathedrals, the firms in which both channels can be seen at scale, technology by technology, and the panel's parents hold 91.2% of all firm-attributed GitHub activity in the linked universe. In the patent bridge, 1,948,899 granted patents of 2015–2024 are attributed to 13,651 parents, and repositories to 2,754 parents. Activity the link misses is not measured: if it sits in the firm's patent technologies more often than the attributed code does, the measured distance overstates the separation, and if less often, it understates it.

The public code. The code comes from GitHub. A harvest of 160,063 numeric GitHub repository identities resolves, at the 2024 parent, to 209,351 attributed repositories. These are the public repositories under the GitHub accounts that a fixed rule links to a parent, so what is measured is what the firm releases as a firm. Not all of it carries an open-source license: by GitHub's license field, read in spring 2026, 58.2% of the eligible repositories carry one, 27.1% carry no detected license and 14.2% a proprietary or custom one, and the more active repositories carry one more often (70% in the most active third, 45.4% in the least). Appendix B.3 re-reads the distance and its excess on the repositories with an open-source license alone. Four kinds of a firm's open-source work fall outside the measure: projects hosted by foundations; repositories transferred to a foundation within the window, which leave the firm's accounts with their whole history; code that comes with an acquisition, which shifts who contributes to the acquired projects (Chen et al., 2022, p. 867); and contributions to repositories the firm does not own. A repository's activity is its dated events in the GitHub Archive, which the held files cover from 2015; the events weight each repository and date its entry. Beyond these exclusions, repositories can move to another account or vanish from GitHub. In a stratified sample of 130 labeled repositories, all 9 that moved to another account moved after their last recorded event under the firm. The labeled repositories that vanished held 0.02% of the labeled repositories' events, weighted to the sample's frame.

The labeler. The model is glm-5.3-flash. The production labels carry two provider names of that one model family, `stealth/ox-alpha` and `z-ai/glm-5.3-flash`, served through OpenRouter at temperature 0 with at most 900 output tokens and low reasoning effort, on the CPC title list of

August 2026; only the first, served name has stopped answering, and the model can still be called under the second, which the reproducibility re-runs use (Appendix C).

The labels' limits. The labels' validation and the run that measured their reproducibility are in Appendix C. As Dell and Rambachan (2026) argue, a measurement built with a language model needs a validation sample drawn from the data it measures; here it is agreement with automated references and a one-coder pilot, and the risk that the model saw the repositories in training is checked by masking (Appendix C). The labels are one retrospective classification pass: they read each public repository's current text and apply it to the repository's whole event history, so the activity is dated but what a repository's code is about is not; relabeling dated snapshots would date it (section 6). The EPO's own text categorizer, trained on examiners' classifications, accepts at most 5,000 characters per request and would need 21 days for the corpus, so it checks the labels rather than producing them. Agreement is measured against the EPO's categorizer on 500 repositories and on a sample of 800 classified through the categorizer's web page, where the production labels share a main group with the categorizer on 76.4% of the items both instruments labeled, and against a 50-item pilot that the author coded blind, one coder's agreement. Label sets change between runs while the portfolio-level objects stay stable, so the paper leans on those objects. A second model's labels, a complete second set given the same prompt and menu, leave the facts in place: the event-weight mean distance rises by 0.046, the two distances correlate at 0.814 across firm-years, and the excess over the randomization benchmark barely moves. That is robustness to which model labels the repositories, not evidence that either labels them correctly. The placement statements rest on one assumption: that label errors are unrelated to the firm's patent technologies and to the channel. Should it fail, a human-coded probability sample of the labels would let the downstream statistics be corrected, not only scored (Wang et al., 2020; Egami et al., 2023; Angelopoulos et al., 2023): treating machine-generated variables as data biases estimates and invalidates inference (Battaglia et al., 2025), and corrections exist for designs, like this one, that aggregate predicted labels into group-level variables (Qiao and Huang, 2025, pp. 29–31). The headline facts are not corrected for label error: without a gold sample, none of these corrections can be applied.

C.4 One repository, end to end

The model reads a repository's name, description, topics, programming language and README, with long READMEs cut (Appendix C). It answers in two steps: first up to three CPC subclasses, then the main groups inside them, judging the repository by what it does rather than what it is written in. For Microsoft's code editor, `microsoft/vscode`, it returns two main groups, `G06F 8/00` (arrangements for software engineering) and `G06F 9/00` (arrangements for program control); the EPO's own text categorizer, given the same repository, returns `G06F 8/00`. Each returned main group takes an equal share of the repository's events, here a half each.

C.5 The populations

Five populations carry the analysis, and every exhibit names the one it uses (Table C5). The firm-year panel holds the parents with patents and linked code in a five-year window; the comparison sample is its part with defined counts over the preceding window, on which the references of the firm's own excess run. The firm-years with an industry group carry the comparison with the firm's industry (section 5.5). The universe of patenting firms, with and without linked code, compares the two kinds of firm within an industry-year, and the firms at risk of beginning to publish code carry the entry hazard. The firm × technology × year grid built on the panel holds 12,061,500 rows; each row is a *cell*: one firm in one technology (one CPC main group) in one year. The grid is balanced over a fixed set of main groups, so its zeros are data; every statement about the share of a firm's open-source mass outside its patent technologies is computed on the full portfolios, not on the grid. Every population inherits the properties of the labels and the links it is built from.

Table C5: The populations

Population	Size	What it holds	Where it is used
The firm-year panel	3,291 firm-years of 748 parents (2019–2024)	Parents with patents and linked code in the five-year window ending in the year	The distance's level; what firms open (sections 4.3 and 3.4)
The comparison sample	2,472 firm-years of 595 firms	The panel's firm-years with defined counts over the preceding window	The firm's own excess and the worlds that reproduce it (sections 4.2 to 4.4)
The firm-years with an industry group	2,450 firm-years of 594 firms	The comparison sample's firm-years whose industry is found in the year before	The firm and its industry (section 5.5; Appendix F)
Patenting firms with and without linked code	17,755 firm-years of 4,087 firms in 661 industry-years; with linked code, 3,561 firm-years of 787 firms	Compustat North America firms with a patent granted 2015–2024 in the patent crosswalk, in fiscal years 2020–2024, with an industry group in the year before	Firms with and without linked code (Appendix F.6)
The firms at risk of beginning to publish code	4,333 firms, one spell each, with 317 first entries; the hazards use the 3,382 spells, with 214 first entries, that leave a year at risk after entry and carry complete covariates	Listed patenting firms observed before their first linked code	The entry hazard (Appendix E.8)

The firm-year panel, the comparison sample, the firm-years with an industry group and the universe of patenting firms, and the firms at risk, are on corrected event dates. The universe of patenting firms counts firms by whether the crosswalk links them to code, not by the year their code appears.

D The facts at a glance

Table D1 gathers the answers to the paper's first question, the measurement, each with what it measures, its reference and its sample, in the order of sections 5 and 5.5. The answers to the second question, the readings against the signs cited sources state, are gathered in Tables E5 and E6.

Table D1: The measurement's answers at a glance

Result	What is measured	Observed	Reference	Sample	What it shows
A. Separation beyond the randomization benchmark, fixed contents and technology norms (sections 5 and 4.4)					
The distance and the benchmark distance	The distance's mean and median; firm-years that share no main group	Mean 0.811, median 0.851; 209 firm-years share none. Comparison sample: 0.7645 (equal), 0.8055 (event)	Randomization benchmark: 0.434 (equal), 0.5719 (event)	Firm-year panel, 3,291 firm-years of 748 parents; comparison sample, 2,472 firm-years of 595 firms	A level alone has no scale; observed portfolios are further apart than the randomization benchmark makes them, and the excess measures the difference
What firms open	Each technology's share of the open-source mass and of the patent mass, pooled over firms	Top four main groups 60.4% of the open-source mass, G06F 8/00 alone 27.61%; sections A–F 0.6%	The same firms' patents: largest main group 3.82%; sections A–F 21.02%	Firm-year panel, 748 parents	Code concentrates in software engineering; patents spread into sections where code is nearly absent
The excess	The observed distance minus the benchmark distance; the share of firm-years where it is positive	0.3304 (equal), 0.2336 (event); positive in 99.4% and 96.5%	Fixed-portfolio world: 0.3336 (equal), 0.2373 (event)	Comparison sample; the world's 40 replicates	More separated than the specified reference; compatible with fixed portfolios
The excess by part	The excess split into code outside the firm's patent technologies and mismatch inside it	Outside 0.2081, inside 0.1223 (equal); outside 0.2001, inside 0.0335 (event)	Each part's own benchmark. Fixed-portfolio world: outside 0.2039, inside 0.1296 (equal); 0.2037, 0.0336 (event)	Comparison sample (Table A1)	With each firm-year counted once most of the mean excess lies outside; weighted by code mass most lies inside (18.12% equal, 6.14% event); firm-year means: 63% (equal), 85.7% (event); both parts compatible with fixed portfolios
Technology norms in the data	The nested references, drawn with the randomization benchmark's device: technology norms, then technology-by-industry norms	Observed 0.7645 (equal), 0.8055 (event); randomization benchmark 0.434, 0.5719; technology norms 0.6035, 0.6864; technology-by-industry norms 0.6048, 0.688	Comparison sample (section 4.4)	Technology norms reproduce 51.3% and 49% of the excess. Industry norms add a further 0.4% and 0.7%, each with an interval that includes zero. The remainder is 48.3% and 50.3%	
Open source follows patenting (Appendix B.5)	The probability of open-source activity in a main group with lagged patent presence; within parent, repositories inside the parent's patented groups against those outside	0.235 points higher. Inside minus outside: a new participant 7.5 points, maintained 6.5 points, retention 0.6 points	Base rate 0.5062%. The same parent's outside repositories; Holm $p < 0.05$, < 0.05 and 0.43	The firm–technology–year panel; the parents' repositories first active 2015–2022	An association within the firm–technology grid and within the parent
The GitHub classes	D_{GH}, the distance on the classes that carry open-source mass; the non-GitHub share ψ_{it}	D_{GH} 0.7998; mean ψ_{it} 0.1758	The full distance: 0.8106	3,278 firm-years with $\psi_{it} < 1$	Setting aside the other classes moves the mean distance little
Concentration and overlap (Appendix B.4)	Firm-years where open source is the more concentrated channel; the distinct-unit HHI gap; firm-years that share a main group	More concentrated in 84% (without duplicated patents, 83.8%); gap 0.0519; overlap 95.3%	Benchmark: gap 0.0026; overlap 99.8%	The GitHub-class firm-years (Table B9)	Open source is the more concentrated channel and overlap is below its level under random assignment; part of the gap is classification, since a patent carries more main groups than a repository
The count slope (section 5.2)	The slope's excess over the randomization benchmark, firm and year effects	0.0516 (event)	Fixed-portfolio world: 0.0463	Comparison sample (Table B2)	Firm-years with more repositories are more separated than the reference; fixed

Table D2: The measurement's answers at a glance, continued

Result	What is measured	Observed	Reference	Sample	What it shows
B. The firm and its industry (section 5.5)					
The six distances (Appendix H)	Each distance's mean excess over its declared reference, read by the exact share test	The firm's code to its industry's code 0.0738; its patents to its industry's patents 0.2989	The industry-pool benchmark; the within-industry benchmark for the portfolios' own separation	2,450 firm-years with an industry group (Table G10)	In each channel the firm's portfolio sits further from both pools than same-size draws; the firm's own two portfolios: not distinguishable from the industry-pool benchmark by the share test, about half the firm-years above it and, on average under event weights, slightly less separated than its draws
With and without linked code (section F.7)	Difference in excesses over the industry-pool benchmark, within industry-years	Patents to the industry's code -0.024; to its patents 0.087	The industry-pool benchmark; Holm $p < 0.001$	17,755 firm-years (Table G8)	Less separated from the industry's code, more from its patents; the raw patent-to-patent comparison points the other way, with no bound on the reversal's sign
C. What moves with the distances (Appendix B.8)					
The firm's own distance and its counts	The coefficient on log repositories within the firm (firm and year effects)	Excess 0.032; observed -0.014	The randomization benchmark	The common sample (Table 7)	The excess rises with the count and the observed distance moves within two standard errors: the reference moves with the count

Event weights; corrected event dates. The firm's industry is its NAICS group in the year before, the firm left out.

E Tests without a finding

This appendix holds the tests of Appendix E.8 whose predictions the data do not bear out at the paper's references: whether code outside firms' patent technologies is more common where patents are harder to get, and the same split run on the firm's patents (E.1); the covariates that the other reasons for opening point to in the entry test, none of which survives Holm's adjustment (E.2), with the fading of the software association; the timing predictions that concern rivals (E.3); and the placement of a firm's code near the earlier code of firms in its industry, with its timing companion (E.4 and E.5). E.6 gives the table of the community contrast within a parent that Appendix E.8 reads, and E.7 the shares that describe the bazaar side. Every test here uses corrected event dates, apart from the twelve-case pilot's counts in E.2, which are labeled.

E.1 The norm: eligibility exposure and the same split on patents

What every holder of a technology does sits in its norm (Proposition 2, part (ii)), so a reason that works through the technology can appear only through technology-level measures. Patent protection is such a reason: where firms choose between patents and other means of protection by how well patents protect in a technology (Cohen et al., 2000), technologies harder to patent should carry more open-source code outside firms' patent technologies, common across firms. Two tests read the technology side. The first asks whether the part of code outside firms' patent technologies that is common to a technology's holders rises with the technology's exposure to patent-eligibility rejections. The second runs the same split on the firm's patents and asks whether the two channels rank technologies the same way.

The exposure is the share of first rejections that carry a §101 rejection, from the USPTO's office action data (Lu et al., 2017), which hold rejections only, measured per technology before the 2018–2019 examination change. Toole and Pairolero (2020a), who follow §101 rejections after *Alice*, count all first actions, allowances included. The paper's mapping of that reason gives a *positive* loading on the mean outside rate across the firms that hold a technology; no cited source states the sign for code, so the loading is a description, not a reading of a stated sign. The loading is estimated on the technology-year means of the outside indicator against the same object-level reference as Appendix B.6, on the same draws.

The loading lies below its reference, on the side opposite to that mapping. The observed loading is 0.0549 against a reference mean of 0.1081 (band [0.0856, 0.1284]), and every one of the 200 draws lies above it (Table E1). The reference's own positive loading is mechanical — firms patent less in technologies with high §101 exposure, so randomly reallocated activity lands outside their patent technologies more often there — and a positive loading beyond random assignment would have to sit above that level; the data sit below it. The technology structure of code outside firms' patent technologies is not accounted for by eligibility exposure beyond what random placement of each firm's own activity produces, on either set of applications or on the exposure measured over the earlier period alone. The reading is a null about one measure: exposure is measured before the

examination change against placement years after it, and whether exposure tracks examination outcomes remains to be shown. A second exposure measure that keeps only eligibility rejections, dropping the statutory double-patenting, utility and inventorship rejections that the office action documentation shows the main measure also counts (Lu et al., 2017), is designed for this loading and is not yet run.

Table E1: Exposure to eligibility rejections against the object-level reference

Statistic	Observed	Reference mean	Reference band	Holm p
Loading on eligibility exposure	0.0549	0.1081	[0.0856, 0.1284]	—

The loading of technology-year mean outside rates on the share of first rejections with a §101 rejection, on 2,160 technology-years, against the reference of Table B16; the band is the draws' 2.5th to 97.5th percentiles, and every one of the 200 draws lies above the observed loading.

The second test runs the same split on the firm's patents, asking whether the technologies a firm patents in outside its open-source technologies rank technologies the way the technologies it opens code in outside its patent technologies do. It is descriptive by construction: no reason for opening signs its statistics relative to the reference without a stated configuration of how the channels respond, so this test cannot falsify patent protection. Read with that qualification: the raw dispersions sit below their reference bands (across firms 0.073 against 0.0932, Holm p 0.0299; across technologies 0.0072 against 0.0458), but this mostly restates a level difference — the observed patent-side outside rate averages 0.8602 against about 0.6681 in the draws, because real open-source mass concentrates in 528 firm-technology cells while a draw spreads it over about 2,524, and a rate near one has little room to vary. Scaled by each assignment's own ceiling on the rate, the dispersion across firms reads *above* its band (0.6068 against 0.4203) and the dispersion across technologies below (0.0601 against 0.2065): accounting for the level, firms' patent-side outside rates differ more than the reference produces while technology-year patterns are more uniform than under random assignment. The agreement between the two channels' rankings matches the reference's own (0.2312 against 0.3256), and on a fixed set of technologies it sits below its band (0.2312 against [0.29, 0.4346], p 0.01). The two channels are asymmetric: the open-source dispersion across firms exceeded its band (Appendix B.6), and the patent-side dispersion exceeds it only after the level is accounted for.

How the tests read. Both tests are descriptions. The loading on eligibility exposure sits below the reference's own mean, below every draw, on the side opposite to the positive loading that the paper's mapping of the choice of protection gives. The same split on the firm's patents reads descriptively. Eligibility exposure is one measure of one reason, and its null leaves the technology structure unnamed rather than unexplained: other technology-level measures, the eligibility-only exposure among them, remain open.

E.2 Entry: the covariates of the other reasons for opening

The entry design. The test needs the attribution of repositories to firms but no patent classes for code, so it stands whatever the label validation concludes. The firms at risk are 4,333 listed

patenting firms not yet publishing when observation begins, each entering the year after its first patent grant, one spell each, with 317 first entries on corrected event dates. The hazards are estimated on the 3,382 spells, with 214 first entries, that leave a year at risk after entry and carry complete covariates: a firm whose first publication falls in its first year at risk enters no hazard model. The event is a firm's first observed activity in the GitHub Archive on a repository linked to it; in a twelve-case pilot, drawn on the event dates as recorded, it matches the repository's creation in 7 of nine repository cases. The estimator is a single-spell Cox hazard (Cox, 1972, §3, eq. 9, p. 189) on calendar years with Efron ties (Efron, 1977, §6, eq. 6.7, p. 563) and robust standard errors (Lin and Wei, 1989, Theorem 2.1, p. 1075). Software-heavy means G06F, G06Q or H04L grant weight, a proxy, since the CPC has no software class.

The other question about openness is whether the firm-level covariates that other reasons for opening point to separate entrants from non-entrants: gross margin, which a protected profit center makes a reason to open in Won (2026b) and a reason not to in Chen (2025); R&D intensity, which goes against opening in Chen (2025); a software type of business, signed by the empirical work on selling complements (Fosfuri et al., 2008, Table 5, p. 300) and its theory (Lerner and Tirole, 2005); and rivals' earlier entry, positive under defensive publication and negative where firms keep knowledge from rivals. On the hazard carrying all four (984 spells, 103 events; software-heavy 4.43 and size 1.86 carried alongside):

Covariate	HR	95% CI	Holm p
Gross margin, per rank SD	1.19	[0.94, 1.51]	0.416
R&D intensity, per rank SD	1.01	[0.74, 1.38]	0.935
Software type (industry rules)	2.49	[1.11, 5.60]	0.109
Rivals' earlier entry, per SD	0.88	[0.71, 1.10]	0.525

None of the four covariates survives Holm's adjustment (Holm, 1979). The software type of business comes closest, at a hazard ratio of 2.49 (95 percent interval 1.11 to 5.60, p 0.027 before and 0.109 after the adjustment). The gross-margin estimate, at 1.19 with an interval of [0.94, 1.51], does not survive correction, and it is a description, since no held record says whether a firm's gross margin measures the profit-center condition of Won (2026a, Proposition 1, p. 8), so no source's sign is read on it. R&D intensity is null, so neither Chen (2025)'s capability sign nor its alternatives is supported at this power. The software-type coefficient points in the direction of the selling-complements prediction of Fosfuri et al. (2008, p. 300, Table 5), but it lies below the design's minimum detectable ratio of 4.117, against a per-standard-deviation detectable ratio of 1.318, so the test has little power. The software-patent prediction is read by the software-stock association of Appendix E.8, where it stands. Rivals' earlier entry shows no association. Manski (1993, abstract) states the reflection problem. Without prior information on the composition of reference groups, a researcher cannot infer whether a group's average behavior influences its members. In this paper's reading, the problem applies here because rivals' entry and a firm's own entry can share an environment, and only rivals among the firms the paper links to repositories are observable. One selection fact stays

on the record beside these estimates: firms with no reported R&D enter at about a third the rate of reporters (the missing-R&D indicator at 0.35), a composition property of the firms at risk.

The fading, and its caveat. The software association of Table E4 weakens over the window. With time interactions centered at 2020, the software ratio at 2020 is 5.11 (95% CI 3.59–7.26) with a per-year multiplier of 0.843 (p 0.037); the calendar windows read 8.15 in 2016–2019 against 2.88 in 2020–2024. When firms differ in their propensity to open in ways the hazard does not see, the likeliest openers leave the risk set first, and hazard ratios shrink over calendar time even when nothing about the association changes: the depletion of susceptibles. On the stacked windows the software ratio is 6.42 early with a software-by-late multiplier of 0.65 (p 0.225) without frailty; with firm-level gamma shared frailty (θ 18.18) the conditional ratio is 12.16 and the multiplier 2.56 (p 0.054). Under the rule declared in advance the fading does not survive frailty, so it reads as consistent with the depletion of susceptibles. The association holds under re-timing by repository creation year (5.73 with a qualifying-license rule, 5.15 without it) and without the 25,423 patents attached to two or more firms at risk, on the hazard's grant table, which includes grants before 2015 (5.07). The twelve-case pilot that traced the event by hand, drawn on the event dates as recorded, also records no detected license in 2 of nine repository cases, a lag between creation and first observed activity in 1, and backdating near the end of the window in 1.

E.3 Order and timing relative to rivals

The defensive-publication predictions of Appendix E.8 concern rivals: a firm that trails its rival in a technology publishes as the rival nears a patent (Bar, 2006, Proposition 3, p. 237); (Parchomovsky, 2000, p. 927). Other models do not turn on leading or trailing (Baker and Mezzetti, 2005, p. 183); (Johnson, 2014, pp. 15–16), and in Gill (2008, p. 733) a leading innovator may disclose. Reading them needs each firm's position against its rivals in each technology. The held records give one: every sample firm's first filing year in every main group it holds. A firm counts as a laggard in a cell when another sample firm's first filing there falls strictly before its own, and as a leader otherwise; only earlier filings enter the rule. The coverage decides against the contrast: of the 3,330 primary cells, 3,327 are laggards and 3 are leaders, with most earlier rival filings 12 to 20 years before the firm's own first filing and a rival filing within five years before it in 59 cells only. That lead is measured on the patents granted in 2015–2024, which reach back only through filings that waited long for their grant: it is set by outliers, while the laggard verdict itself is robust. The population's own censoring, a first filing in 2016 or later, selects late entrants into cells whose rivals filed long before, so the leader side of the contrast is empty.

Two further checks reach the same verdict. Comparing firms that entered the same main group in the same three-year window supplies the coverage: 3,323 of the primary cells sit in usable groups of such entrants, 888 of them first-year entrants. But measures of a firm's position at entry, its stock of related patents and its recent specialization in related technologies, fixed at a common 2014 baseline through the patent-to-firm match of Kogan et al. (2017), find no support on this

population: a high stock position covers 24 of 3,067 entrant cells (17 firms), the mass sitting in the bottom decile, because the entrants are almost without exception the low-position side of the incumbents in their cells. The defensive-publication predictions therefore cannot be identified on this design; a timing contrast within groups of firms that entered together is the open option.

E.4 Code near same-industry firms' earlier code

Does a firm's code in a technology lean toward the code that firms in its own two-digit SIC industry published there earlier, more than the randomization benchmark produces? The test uses the co-location scores of Appendix E.8, on the same technologies and the same 200 reference draws, with every exposure rebuilt on each draw, over 2,450 firm-years. The outcome is the firm's open-source presence in a technology; a companion reads, where both channels are present, the mix of open-source events against patent grants, of which only the covariance is read. The exposure is the share of same-industry firms' code in the technology over the trailing five years. Two companions use other peers: product-market rivals in the sense of Hoberg and Phillips (2016), on the firms that classification covers, and the ten nearest patent neighbors. No reason signs the association, so every test is two-sided.

The co-location is beyond the reference: the same-industry presence score is 0.02105 against a reference mean of 0.00447 (band [0.00395, 0.00511]; Holm p 0.0199), and the mix score is 0.01206 against 0.00459 (band [0.00374, 0.00533]); the product-market companion (0.01506) and the neighbor companion (0.03106) agree. Two checks declared in advance qualify it. In a world with no peer term, where a firm's presence in a technology is drawn at the rate of the other firms of its industry in that technology, all 10 worlds put the presence score beyond their own bands in the observed direction: a taste for technologies shared within an industry produces the excess with no peer term at all. And with firm-year and technology-year effects absorbed, and the firm's own lagged code share and patent position controlled, the association sits inside its band (-0.0263, band [-0.1033, 0.0707], p 0.796), while the firm's own earlier code share is large and beyond its band (0.3029, p 0.0199). The firm's code lies nearer its industry's earlier code than the randomization benchmark produces, as industry-level tastes for technologies would place it; the test gives no evidence of a peer term.

E.5 Timing after same-industry firms' first code

Does a firm's code in a technology change in the years after firms of its own two-digit SIC industry first publish code there? This timing companion to the placement test above was declared before any output. Its counterfactuals are imputed from unexposed rows, in the manner of Borusyak et al. (2024): a model with three crossed fixed effects (firm × technology, firm × year, technology × year) is fit on the rows not yet exposed, and its predictions stand in for the exposed rows' counterfactuals.

The declared rule asks that at least 0.60 of the exposed post rows carry complete fixed-effect support before any reading. On corrected event dates, with each firm's entry into a technology dated by its fixed entry clock, only 3,108 of the 12,483 exposed post rows do, a share of 0.249, and none of the 4,221 rows of the 2019 cohort or the 5,115 of the 2020 cohort is estimable under the declared

rule, so the test is not estimable as declared. The program fits each exposed cell's own rows before $t^{\exp} - 4$, with $t^{\exp}$ the cell's exposure year, as the contract's estimator states; its population text gives the own fit rows one more year, to $t^{\exp} - 4$, and the verdict stays as run. The crossed fixed effects, fit on unexposed rows, cannot identify counterfactuals for most exposed firm-technology cells in the technologies where same-industry firms open code, which are the cells the event study needs. This is a property of the design on this population, not a defect of its code, and no reading of timing after same-industry firms' code follows from it. The falsifications that the design's reporting rules name — a placebo on unrelated CPC groups, a gradient in similarity, and a contrast with technology neighbors that are not product-market rivals — were not run, and their absence is not evidence about any of them.

E.6 The community within a parent

This subsection gives the contrast within a parent that Appendix E.8 reads. Within the same parent and first-activity year, a repository is *inside* when at least half of its label mass lies in main groups where the parent was granted a patent in the five years to its first-activity year; each row regresses an outcome on the inside indicator with parent and cohort effects, standard errors clustered by parent, Holm's adjustment across the four rows.

Table E2: Inside against outside the parent's patented technologies, within parent

Outcome	Inside minus outside (points)	SE	Holm p
A participant new to the repository within two years	7.5	0.9	< 0.05
One with no prior event on the parent's organizations	8.1	1.0	< 0.05
A push or a release in the second year	6.5	1.1	< 0.05
New participants still active a year later	0.6	0.8	0.43
Repositories (inside)	110,500 (87,670), of 724 parents		

Linear probability; parent and first-activity-year effects; standard errors clustered by parent. A new participant is an actor new to the repository; the second row's proxy for people from outside the firm is an actor with no prior event on the parent's linked organizations, a statement about event histories, not employment. Corrected event dates.

E.7 The bazaar side

This subsection describes what corporate repositories draw, the community side of opening, with no reference and no test. Of the 119,167 repositories that the sample's parents first make active in 2015–2022, 50.7 percent draw a participant new to the repository within two years, and 38.6 percent draw one with no prior event on the parent's organizations, a declared proxy for people from outside the firm. Few stay: 11.7 percent of the new participants are still active on the repository a year later. Under half of the repositories, 40.3 percent, receive a push or a release in their second year, and the top 1 percent of repositories hold 50.5 percent of the arrivals. These shares use corrected event dates. They describe the community side and measure no gain to the firm from opening; on the gains, Singh et al. (2026, p. 1) relate high-tech firms' adoption of open-source product development

to higher gross margins.

E.8 The stated reasons: every reading

The measured differences cannot yet tell the literature's reasons for opening apart. Of the nine stated signs that a statistic reads, five match, one does not match and three are not decided (Table E3). One match reads an object near its reason, the co-location of a firm's code with its nearest technology neighbors' patents. The other four say less. Any technology mix specific to the firm gives the sign of IP modularity, read in module terms by this paper's mapping, and the sign of involuntary spillovers. Technology position predicts the sign of entry, and the order of the channels matches an order stated for analogues, not for code. Grouped by whom a reason concerns (rivals, the community, or customers and users), the rivals' reasons bear on four rows and the community's on one. The customers-and-users reason has no test. Four rows, IP modularity, the order, the activity around a filing and entry, read statements that no such group makes. For IP modularity, the test of its condition with outside participants is not consistent with the recommended separation (section 5.4). Four further rows are described rather than read, and three have no statistic (Table E6). Appendix I says where and when each sheet of stated signs was fixed. The six distances' sheet was fixed before those comparisons were computed, and the sheet for the paper's own tests after those tests had run.

The reading is *matches*, *does not match* or *not decided* against the stated sign, whichever way the data point. A row enters when a cited source states a sign for the quantity the row reads, or for an analogue that the row names as one, not a mechanism only. A statistic coarser than the statement reads *not decided* at the design's resolution, a row that no statistic read is *not read*, and a statement that no statistic the data allow could read, such as a timing statement that needs rivals' dates, is *not identifiable*, a verdict of its own. "Matches" says that the data are consistent with a stated sign, and nothing about a motive; several reasons can share a sign, so no row singles out a reason.

Close rivals. A firm's code outside its patent technologies is less co-located with its nearest patent neighbors' patent technologies than the randomization benchmark produces. Each score multiplies the share of the firm's outside code in each technology by a peer group's patent presence there, centered within the firm-year, and sums, against the randomization benchmark with every input built from patents rebuilt on each draw. A score measures co-location, not distance: more code where the peers patent does not fix the sign of a distance to them (Lemma 13). Each score carries its own Holm-adjusted p (Holm, 1979); Appendix B.7 defines the three peer groups and shows that dropping shared patents changes no decision. The score with the ten nearest patent neighbors is -0.00596 against a reference mean of -0.00427 (Holm p 0.0199). Within technology-years, a weighted regression of the firm's outside placement in a technology on its ten nearest neighbors' patent share there, with firm-year and technology-year effects and the firm's own patent position as a control, gives a slope of -2.3827 per unit of the neighbors' patent share, which runs from zero to one. Its difference from its mean under the randomization benchmark is -0.219, and the slope

Table E3: The stated reasons: every reading

Reason; the source's stated sign	Statistic and reference	Estimate	Any firm-specific mix or technology position gives the sign?	Reading
What is revealed and what is protected				
IP modularity. Henkel et al. (2013, p. 72); West (2003, p. 1279): revealed parts in open modules or layers, apart from what the firm protects, recommended under conditions	The excess of the firm's open-source code over its patents against the randomization benchmark, and its share in main groups where the firm holds no patent (sections 4.3 and 4.3)	Excess 0.234 (equal weights 0.330); outside share 85.7% (97.3% under another model's labels; equal weights 63%), each firm-year counted once; 6.14% with firm-years weighted by their events	Yes: any firm-specific mix	Matches for the excess; not consistent for its condition (section 5.4)
Rivals				
Close rivals. Llanes (2019, Prop. 5); Harhoff et al. (2003): less openness where rivals use more of what is disclosed	Co-location of the firm's outside code with its ten nearest neighbors' patent technologies, against the randomization benchmark; within technology-years (Appendix E.8)	−0.00596 against −0.00427 (Holm p 0.0199); within technology-years, the slope's difference from the randomization benchmark is −0.219	No	Matches
Defensive publication. Bar (2006); Parchomovsky (2000): the firm behind publishes, in race terms	The co-location with peers that grow where the firm lags, the same reference	Across technology-years 0.00342 against −0.00205; within, −0.777 against 0.003	No	Not decided
Commoditizing a complement. Won (2026a, Prop. 1, p. 8): release in the complement market, under conditions	The co-location with complementors' patent technologies, the same reference	Inside its band (Holm p 0.532)	No	Not decided
Involuntary spillovers. Kamien and Zang (2000): firm-specific research approaches	The firm's patents to its industry's patents, against the industry-pool benchmark (Appendix H)	0.2989 [0.2868, 0.311] (Holm p 0.02)	Yes: any firm-specific mix	Matches
The community				
The community. Lin and Rai (2024, p. 662), summarizing prior work, and Wen et al. (2013), after an enforcement action: less participation in projects related to patented inventions	New participants inside a parent's patented technologies against its outside repositories, within parent (Appendix E.8)	7.5 points higher (Holm $p < 0.05$)	No	Does not match
Order, timing and entry: statements no group of reasons makes				
The order of the channels. Ganglmair and Kann (2025a, slide 20); Argente et al. (2025): an analogue comes after the patents	The share of cells where code comes first, against the within-firm permutation of activity profiles (Appendix E.8)	41.56% against a band of 47.8% to 50% (Holm p 0.012)	No	Matches the analogues' order
Activity around a filing. Petralia (2025): a spike in the month before a filing	The share of a cell's code activity in the two years ending at its first filing, the same reference	13.32% against 13.73% to 15.04% (Holm p 0.016); by grant date, inside its band	No	Not decided at the design's resolution
Entry. Fosfuri et al. (2008); Coleman et al. (2025): more releases and adoption with more software patents	The hazard of first publication among listed patenting firms (Appendix E.8)	Hazard ratio 5.03 [3.60, 7.04]; 4.47 beside the technology norm	Yes: technology position	Matches

Every reading of a sign that a cited source states for the quantity the row reads, or for a named analogue. A reading states a test's verdict against the stated sign, never a motive; several reasons share signs, so no row singles out a reason. The introduction's customers-and-users reason has no row. Intervals in brackets are 95 percent intervals, clustered by firm. Table E5 gives each source's statement in full, and Table E6 the rows that are described or not read.

lies just below the lower edge of the band of the benchmark's draws, which runs from −2.3818 to −1.9631 (declared one-sided $p = 0.03$). The benchmark's draws of the score have a standard deviation of 0.000115, and those of the slope 0.1065. The score's difference from the reference mean is −14.7 of its draws' standard deviations, and the slope's is −2.05 of its own, both on the side of less co-location with the close technology peers than the object-level reference gives. The reading below rests on the score. Llanes (2019, Proposition 5, p. 288) states that firms open less when rivals can use more of what they disclose, and Harhoff et al. (2003, p. 1762) that stronger competition makes free revealing less likely. The reading *matches* their sign; measuring it by co-location is the paper's mapping.

Among the ten nearest patent neighbors, those that the text-based network industry classification (TNIC-3) of Hoberg and Phillips (2016) lists in the year before as the firm's product-market competitors can be set beside unlisted neighbors matched to them one to one on technological proximity. The comparison uses the 975 of 2,450 firm-years, of 291 firms, whose ten neighbors include at least one listed and one unlisted neighbor. Its reference is the randomization benchmark, with the neighbors and the split rebuilt on each draw, and the three scores form one Holm family. With the listed competitors, the score is −0.00617 against a reference mean of −0.00407 (Holm p 0.0149). With the matched unlisted neighbors, it is −0.00609 against −0.00395 (Holm p 0.0149). Their difference, −0.00008, lies inside the band of its draws, from −0.00039 to 0.00013 (Holm p 0.6169). The lower co-location is thus about as large with technology neighbors that are not product-market competitors as with those that are. This is an association in technology space, and the data do not single out product-market rivalry over technological resemblance.

Defensive publication. Bar (2006, p. 237) and Parchomovsky (2000, p. 946) state the sign in race terms: the firm behind publishes in the race its rival is close to winning; reading that as code near the patents of peers that grow where the firm lags is the paper's mapping. Across technology-years the score lies in that direction, 0.00342 against −0.00205 (Holm p 0.0199); within technology-years it runs the other way, −0.777 against 0.003 (two-sided p 0.00995). The sign is conditional in Bar (2006, §4.1, p. 239) and in Baker and Mezzetti (2005, Propositions 1 and 2). With one statistic in the stated direction and one against it, the reading is *not decided*. The timing statement is not identifiable without rivals' dates (Appendix E.3).

Commoditizing a complement. Won (2026a, Proposition 1, p. 8) states, in market terms and under conditions, that a firm with large profit-center rents and a strong cross-market demand link releases an open-source asset in its complement market. The complementors' score sits inside its band (Holm p 0.532), on thin support; reading that market as the complementors' patent technologies is the paper's mapping, and the test does not condition on a dominant profit center. The reading is *not decided*. The joint test across the peer groups rejects agreement with the reference (J 11.42) without saying which group departs.

The order of the channels. Code comes first in a technology less often than a permutation of the firm's own activity profiles places it: in the cells where a firm has both channels and a first patent filing from 2016 on, code comes first in 41.56% of cells, below the permutation's 95 percent band of 47.8% to 50% (p 0.004; Holm 0.012), and the comparison holds when patents are dated by grant. No source the paper found states this order for code; two state it for analogues, research disclosures that go to older lines of research (Ganglmair and Kann, 2025a, slide 20) and patent filings followed by product introductions (Argente et al., 2025, abstract; p. 23), and the reading *matches* the analogues' order. Appendix B.7 defines the cells and the permutation (Tables B18 and B19).

Activity around a filing. The mean share of a cell's code activity in the two years ending in its first filing year is 13.32%, below the reference band of 13.73% to 15.04% (p 0.008; Holm 0.016), and the same share with patents dated by grant lies inside its band (p 0.37). Petralia (2025, p. 9) states a spike in the month before a patent's filing, finer than the design's annual dates, so the reading is *not decided* at the design's resolution, and no contradiction of a month-level statement.

Entry. Firms whose patent stocks are software-heavy begin to publish code at a hazard ratio of 5.03, and at 4.47 with the leave-one-firm-out technology norm beside them, the reference that the identification proposition names (Table E4). Among firms that introduced open-source products, a larger stock of software patents goes with more releases (Fosfuri et al., 2008, p. 299), and firms filing more software patents adopt GitHub more often (Coleman et al., 2025): release counts and adoption, where this reads a hazard of first publication beyond the technology norm. The reading *matches* their sign. Technology position predicts the same sign, since the indicator counts patents in the classes where code concentrates. The proportional-hazards tests give p 0.0248 for column 1 and 0.0136 for column 2, and the ratio falls from 8.15 in 2016–2019 to 2.88 in 2020–2024, so column 1's ratio averages over a ratio that falls with time. Appendix E.2 gives the design, the two windows and the fading under firm frailty.

The community within a parent. Within the same parent, repositories inside its patented technologies draw a participant new to them within two years at a share 7.5 points higher than its outside repositories, and are maintained at a share 6.5 points higher (Holm $p < 0.05$ and < 0.05; Appendix Table E2). Lin and Rai (2024, p. 662), summarizing prior work, state that projects built on or related to patented inventions draw less user interest and fewer voluntary contributions; Wen et al. (2013, p. 1131) state it after an enforcement action, for projects overlapping the disputed open-source application, and Bar-Gill and Parchomovsky (2003, pp. 1860–1861) give a mechanism about follow-on inventors. The reading *does not match* the sign that Lin and Rai (2024) summarize.

Involuntary spillovers. Among the predictions stated for the six distances, the involuntary-spillovers row matches. Kamien and Zang's firm-specific research approaches, which limit involuntary spillovers (Kamien and Zang, 2000, abstract; Remark 2), predict the firm's patents further from

Table E4: Which patenting firms start publishing code: hazard ratios

	(1) Own patent classes	(2) With the technology norm
Software-heavy patent stock	5.03	4.47
	[3.60, 7.04]	[3.14, 6.36]
Technology norm, per SD		1.14
		(p 0.033)
Firm spells	3,382	3,381
First entries	214	214
Proportional-hazards test, p	0.0248	0.0136

Single-spell Cox hazard on calendar years, Efron ties, robust standard errors; 95% intervals in brackets. The proportional-hazards test is that of Grambsch and Therneau (1994, §2, eqs. 9–11, p. 517). The firms at risk are listed patenting firms from the year after their first patent grant, the spells that leave a year at risk after entry and carry complete covariates (Appendix E.2); the event is a firm's first observed GitHub Archive activity on a repository linked to it, on corrected event dates. Both models also carry a dominant core business and size (log assets). Levels of assets and sales are computed among US-dollar reporters, the Compustat North America rows reported in US dollars; spells without a US-dollar level drop from the models that use it. Software-heavy: G06F, G06Q or H04L grant weight in the firm's own patent classes. The technology norm is the share of the other holders of the firm's patent classes that already show open-source mass there, so column 2 reads the association beyond the firm's technology position. Appendix Table B20 gives the other covariates, the other event dates and rules, the two windows and firm frailty.

its industry's patents than the industry-pool benchmark places them; they are 0.2989 further on average (Holm p 0.02), a sign any technology mix specific to the firm gives.

The six distances' other rows. The other predictions stated for the six distances are described or not read (Table E6).

E.9 The year a firm starts to publish open-source code

In the year a firm starts to publish open-source code, neither excess of its patents' two distances to its industry is distinguishable from that of other at-risk firms, once the two are tested together.

The test regresses the excess of each of two patent distances on the year the firm starts (Table E7), against at-risk firms that do not start that year. The firms at risk are those of the entry hazard (Table E4): listed patenting firms that have not yet published open-source code. Each is at risk from the year after its first patent grant until and including the year it starts, and a firm starts in the year of its first observed activity on an open-source project linked to it. A firm without open-source code has two of the six distances of section 5.5: one runs from its patents to its industry's open-source code, and the other from its patents to its industry's patents. I measure both on the five years that end the year before. For each distance, the excess is the observed distance minus the benchmark distance, and both distances use the industry-pool benchmark (section 5.5). For these firms, the first window with an excess ends in 2020, so I use the at-risk years 2021 to 2024. They hold 11,753

Table E5: What the measured differences say about the stated reasons: the paper's own tests

The source's statement, its direction for the row	What reads it: statistic, object, reference, population	Reading
IP modularity. Henkel et al. (2013, p. 72): when value co-creators are distributed, numerous and anonymous (Table 1, p. 70), relatively unprotected open modules beside highly protected closed ones are recommended (same in module terms, as a recommendation; main groups in place of modules and the firm's patented technologies in place of its protected modules are [the paper's] mapping; the condition, read with outside participant-years as a proxy, goes with less code outside the patent support, section 5.4; combining weak and strong protection in one module, p. 77, is the opposite arrangement); West (2003, p. 1279): control waived over a platform's commodity layers, kept over the layers that differentiate (same (an analogue: platform layers in place of technologies))	The excess of the firm's open-source code over its patents against the randomization benchmark, and the share of that excess in main groups where the firm holds no patent; panel firm-years (sections 4.3 and 4.3)	Matches for the excess; any firm-specific mix gives the sign; the test of the condition is not consistent with the recommended separation
Close rivals. Llanes (2019, Prop. 5, p. 288): less openness where rivals use more of what is disclosed (same); Harhoff et al. (2003, p. 1762): stronger competition, less free revealing (same)	The co-location of the firm's outside code with its ten nearest patent neighbors' patent technologies, against the randomization benchmark with peers rebuilt on each draw; panel firm-years, within technology-years (Appendix E.8)	Matches
Defensive publication. Bar (2006, p. 237): the follower publishes when the leader is close to the patent (same in race terms; see the note); Parchomovsky (2000, p. 946): trailing firms publish preemptively (same in race terms; see the note)	The co-location with peers that grow where the firm lags, the same reference; across technology-years in the stated direction, within technology-years the other way (Appendix E.8)	Not decided; timing not identifiable
Commoditizing a complement. Won (2026a, Prop. 1, p. 8): release in the complement market, under conditions (same in market terms, under the proposition's conditions; the complementors' patent technologies are [the paper's] mapping)	The co-location with complementors' patent technologies, the same reference; no condition on a dominant profit center (Appendix E.8)	Not decided
The order of the channels. Ganglmair and Kann (2025a, slide 20): disclosures go to older lines (same (an analogue: research disclosures in place of code)); Argente et al. (2025, abstract; p. 23): patent filings followed by product introductions (same (an analogue: product introductions in place of code, in consumer goods))	The share of cells where code comes first, against the within-firm permutation of activity profiles; cells with a first filing from 2016, at annual dates (Appendix E.8)	Matches the analogues' order
Activity around a filing. Petralia (2025, p. 9): contributions spike in the month before a patent's filing (opposite)	The share of a cell's code activity in the two years ending at its first filing, the same reference and cells (Appendix E.8)	Not decided at the design's resolution
Entry. Fosfuri et al. (2008, p. 299): more releases with a larger software-patent stock (same); Coleman et al. (2025): more GitHub adoption with more software patents (same)	The hazard of first publication among listed patenting firms, a software-heavy stock beside the technology norm (Appendix E.8)	Matches
The community. Lin and Rai (2024, p. 662): less user interest and contribution for projects related to patented inventions (opposite (prior research cited, not tested)); Wen et al. (2013, p. 1131): the same after an enforcement action (opposite (after an enforcement action, for projects overlapping the disputed open-source application))	The new-participant share of a parent's repositories inside its patented technologies against its outside repositories, within parent (Appendix E.8)	Does not match

Each source's direction for the row is the literature review's label, in parentheses, word for word. The defensive-publication row's labels in full: for Bar, "same in race terms (the firm behind publishes in the race its rival is close to winning); code in place of publication and peers that grow where the firm lags in place of a race's leader are [the paper's] mapping; no stated direction relative to the firm's own patents"; for Parchomovsky, "same in race terms (trailing firms publish preemptively so that the leader cannot patent); code in place of publication and peers that grow where the firm lags in place of a race's leader are [the paper's] mapping; no stated direction relative to the firm's own patents". A reading states a test's verdict against

Table E6: What the measured differences say about the stated reasons: the six distances

The source's statement	What reads it	Reading
Involuntary spillovers. Kamien and Zang (2000, abstract; Remark 2): firm-specific research approaches	The firm's patents to its industry's patents, against the industry-pool benchmark (Appendix H)	Matches; any firm-specific mix gives the sign
Spillover pools. Cao et al. (2026, Table 4): more new repositories where the firm's patents sit nearer others' patents	Its declared statistic put the distance on the right, withdrawn	Not read
Entrants. Cockburn and MacGarvie (2011, p. 931): some entrants sidestep patent thickets in open source	Panel age on the firm's code to its industry's patents, beyond the randomization benchmark: −0.001623 (0.00116)	Described
A fence. Lerner (1995): firms with high litigation costs patent less where competitors' patents sit	Panel age on the firm's patents to its industry's patents, beyond the industry-pool benchmark: 0.0002376 (0.002419)	Described
A race. Cabral (2002, Lemma 1; Prop. 1): a leader far ahead and its laggard choose differently	The leader indicator on the same distance, beyond the industry-pool benchmark: 0.01771 (0.01775)	Described
Patent strength. Comino and Manenti (2022, Prop. 1; Cor. 2): either way, or nearer where patent reach extends into adjacent technologies	The exposure to the 2019 guidance on the same distance, beyond the industry-pool benchmark, in standard deviations: −0.1963	Described
Persistence. Stuart and Podolny (1996, pp. 27–31): relative positions stable across windows	No statistic	Not read
Ownership. Ziedonis (2004, Table 2); Leten et al. (2016, Table V): by the ownership of the industry's patents	No statistic	Not read

Each reading follows a sheet fixed before the comparisons were computed; the four described rows give a coefficient with standard error, clustered by firm, from a regression of the distance's excess on the firm's variables with industry and year effects, on proxies the paper chose (panel age for entry and litigation costs, the R&D-rank leader for a race lead, the exposure to the 2019 guidance for patent strength). Log assets and the leader are computed among US-dollar reporters.

firm-years of 3,460 firms, with 94 starts. I regress each excess on an indicator for the year the firm starts, with industry-year fixed effects, so each regression compares firm-years of the same industry and year. Column 1 adds the entry hazard's covariates that need no US-dollar level: the firm's earlier patent grants, a software-heavy patent stock and a dominant core. Column 2 also adds log total assets in US dollars, which only the US-dollar firm-years have. I cluster standard errors by firm. Before the run, I declared the two distances one family of tests, and Holm's adjustment (Holm, 1979) corrects their p-values for running two tests.

Once both distances are judged together, the lean of a starting firm's patents toward its industry's open-source code is not distinguishable from zero. For the excess of the distance to the industry's open-source code, column 1 gives a coefficient of -0.011 on the year the firm starts. The coefficient is in the distance's units, the share of the patents' weight that would have to move to other main groups to match the industry's portfolio, and a negative coefficient means a smaller excess. Within an industry-year, the point estimate puts a starting firm's patents nearer the industry's open-source code than other at-risk firms' patents, against the industry-pool benchmark. Its p-value is 0.193 before Holm's adjustment and 0.386 after it. Under equal weights, it keeps its sign (-0.012). The distance to the industry's patents has a coefficient of 0.016, with a Holm-adjusted p-value of 0.386. Neither coefficient is distinguishable from zero at the five percent level after the adjustment. Column 2 gives the same picture for the two distances: -0.014 (Holm-adjusted p-value 0.257) and 0.013 (0.457). In economic terms, the point estimates say that a starting firm's patents lean toward the technologies of its industry's open-source code, more than the industry-pool benchmark places them.

Table E7: The year a firm starts to publish OSS: the excess of its patents over the industry-pool benchmark

	(1) Without log total assets	(2) Adding log total assets, US-dollar firm-years
A. The firm's patents and the industry's OSS		
Year the firm starts, event weights	−0.011*	−0.014
	(0.009)	(0.009)
Holm-adjusted p-value	0.386	0.257
Year the firm starts, equal weights	−0.012*	−0.015
	(0.009)	(0.01)
Firm-years	9,519	8,387
Firms	3,027	2,636
Starts	82	71
B. The firm's patents and the industry's patents		
Year the firm starts	0.016	0.013
	(0.015)	(0.017)
Holm-adjusted p-value	0.386	0.457
Firm-years	9,756	8,572
Firms	3,099	2,693
Starts	82	71

OSS means open-source software. Each coefficient is on an indicator for the year a firm starts to publish open-source code. Each regression puts the excess on the left and absorbs industry-year fixed effects. Standard errors clustered by firm are in parentheses. The firms at risk are the entry hazard's (Table E4): listed patenting firms that have not yet published open-source code. Each is at risk from the year after its first patent grant until and including the year it starts. A firm starts in the year of its first observed activity in the GitHub Archive on an open-source project linked to it. The at-risk years are 2021–2024. A firm-year enters when the firm has patents in the window and an industry in the year before. Industry-years with a single firm-year drop out. Each distance is measured on the five-year window that ends the year before. The industry is the firm's industry in the year before (section 5.5). The industry's portfolios leave the firm out. Event weights count each open-source project by its events, and equal weights count each project once. The distance between the firm's patents and the industry's patents has one weighting, since it holds no open-source code. The industry-pool benchmark deals the technologies of the industry-year's items out again within each channel and each item's start year. Its pool includes the firm's own items. Each portfolio then becomes a draw of its own size from its channel's pool. The benchmark distance is the distance's mean over the benchmark's draws. The excess is the observed distance minus the benchmark distance, in the distance's own units. It is positive when the two portfolios are further apart than the industry-pool benchmark makes them. A negative coefficient means a smaller excess: against the industry-pool benchmark, the starting firm's patents sit nearer the industry's portfolio than other at-risk firms' patents do. Column 1 needs no US-dollar level. It carries three covariates: the log of one plus the firm's patent grants before its first year at risk; an indicator for a software-heavy patent stock (the share of those grants in the main groups of G06F, G06Q or H04L lies above its median among the entry hazard's firms); and an indicator for a dominant core (the largest main group's share of those grants lies above its own median among the entry hazard's firms). Column 2 adds log total assets in the first year at risk, standardized. It exists only for the US-dollar firm-years, those with a Compustat North America row in US dollars. Holm's adjustment (Holm, 1979) runs over the two event-weight coefficients of a column, the family of two tests declared before the run. The equal-weight row lies outside the family and has the event-weight row's firm-years, firms and starts. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from each coefficient's own p-value before Holm's adjustment.

F The six distances under other weights, groups and references, and firms with and without linked code

This appendix reads the six distances among the firm's two portfolios and its industry's (Appendix A.13) under other weights, with other groups in the industry's place and against other references, and compares firms with and without linked code. Each is read against a named reference with the test of Appendix A.14; none is read as a level alone. Except in the last subsection, the frame is the comparison sample: 2,472 firm-years of 595 firms, of which 2,450 have an industry group. An excess is the observed distance minus the reference's mean; a positive excess is further apart than the reference places the firm, a negative one nearer.

F.1 Equal weights

The firm's own distance is not distinguishable from the industry-pool benchmark under the share test, and two descriptions sit beside that reading without changing it (Appendix H). The mean excess is −0.0231 under event weights, with a 95 percent interval over 51 industry clusters from −0.0414 to −0.0048, and −0.0013 under equal weights, from −0.0276 to 0.025; and the pools hold the firm's own objects (section 5.5).

Table G1 reads each distance against its declared reference under equal weights, which count each repository once. Under equal weights, the firm's own two portfolios are further apart than the within-industry benchmark places them, and so are the industry's. Against the industry-pool benchmark the share test does not reject for the firm's own distance. Each of the four distances from the firm to its industry sits further apart than the industry-pool benchmark places it, and leaving the firm's own objects out of the industry's pools leaves the industry's two portfolios further apart than leaving out draws of the same size.

Table G1: The six distances against their declared references, equal weights

Distance	Reference	N	Observed	Ref. mean	Excess	Share above	Holm p
The firm's patents and its code	Within the industry	2,450	0.7651	0.5699	0.1951	0.879	0.04
The industry's patents and its code	Within the industry	2,439	0.6602	0.2467	0.4135	1.0	0.04
The firm's patents and its code	Industry pool	2,450	0.7651	0.7664	−0.0013	0.548	0.576
The firm's code and the industry's code	Industry pool	2,441	0.5076	0.3313	0.1763	0.941	0.04
The firm's code and the industry's patents	Industry pool	2,448	0.8055	0.7405	0.065	0.858	0.04
The firm's patents and the industry's code	Industry pool	2,441	0.7761	0.7164	0.0597	0.717	0.04
The firm's patents and the industry's patents[a]	Industry pool	2,448	0.6784	0.3794	0.2989	0.989	0.02
The industry's patents and its code	Industry pool	2,439	0.6602	0.6558	0.0045	0.615	0.04

Firm-years of the comparison sample with an industry group. "Within the industry" is the within-industry benchmark, which reassigns the industry-year's objects, the firm's included, across channels and owners within start years; "Industry pool" is the industry-pool benchmark, which reassigns them within each channel and start year (Appendix A.14). Observed and reference means are averages over firm-years; the excess is their difference. The share above is the share of firm-years above their reference, pooled over the years as a description; Holm p is the smallest over the five years of the exact share test's p under equal weights, each adjusted by Holm's procedure over the row's tests, its two weightings by five years (the five years alone where only patents enter); below 5 percent, the test rejects in at least one year. For the industry's own distance the firm is left out, so its reference rebuilds the industry's portfolios from each draw without the firm's slots.

[a] The firm's patents and the industry's patents carry no repository weights, so this row is the same under both weightings.

F.2 The industry's leaders as the group

In Table G2 the group is the firms in the top third of the industry-year by R&D rank among US-dollar reporters, with the firm left out. Against the within-industry benchmark, the firm's own two portfolios and the leaders' are further apart than that benchmark places them, under both weightings. Against the industry-pool benchmark, the share test does not reject for the firm's own distance under either weighting; each of the four distances from the firm to the leaders sits further apart than the industry-pool benchmark places it, and leaving the firm's objects out leaves the leaders' two portfolios further apart than leaving out draws of the same size.

Table G2: The industry's leaders as the group: each distance against its declared reference

			Event weights			Equal weights		
Distance	Reference	N	Excess	Share above	Holm p	Excess	Share above	Holm p
The firm's patents and its code	Within	2,450	0.118	0.811	0.04	0.2103	0.894	0.04
The leaders' patents and their code	Within	2,297	0.2841	0.997	0.04	0.3898	1.0	0.04
The firm's patents and its code	Industry pool	2,450	−0.0162	0.529	0.32	0.0053	0.547	0.324
The firm's code and the leaders' code	Industry pool	2,308	0.0818	0.696	0.04	0.184	0.931	0.04
The firm's code and the leaders' patents	Industry pool	2,399	0.0421	0.801	0.04	0.0622	0.842	0.04
The firm's patents and the leaders' code	Industry pool	2,308	0.0792	0.78	0.04	0.0743	0.748	0.04
The firm's patents and the leaders' patents	Industry pool	2,399	0.3132	0.989	0.02		(the same)	
The leaders' patents and their code	Industry pool	2,297	0.027	0.823	0.04	0.0057	0.605	0.04

The leaders are the firms in the top third of the industry-year by R&D rank among US-dollar reporters, with the firm left out. A firm outside the top third forms its own pool with the leaders' objects. "Within" is the within-industry benchmark's leaders; "Industry pool" is the industry-pool benchmark on the leaders' pool. N is the number of firm-years under event weights. Share above and Holm p as in Table G1.

F.3 All other owners, and the firms of other industries

Table G3 reads the four distances from the firm to the portfolios of every other owner, with the firm left out, against the randomization benchmark of the firm's own objects; the portfolios of every other owner are held as observed. Two definitions of all other owners stand side by side: every patent, whoever holds it, with every linked repository counted once; and every firm the crosswalk matches, with each firm's linked code summed. Against both, the firm's code sits nearer the others' code than the randomization benchmark places it, and further from the others' patents; its patents sit further from the others' code and from the others' patents. The firm's share of all owners' portfolios is small on average: 0.0017 of the code under event weights and 0.001 of the patents. By Lemma 14(v), a pool that holds the firm sits nearer its own portfolio by exactly the factor of one minus that share. The distance between the firm's code and the others' code with the firm in the pool, minus the same distance without it, averages −0.0006 under event weights; for its patents the same difference has a slope of −0.573 on the firm's patent share.

The last row reads the firm's patents against the patents of firms in other industries. Under one draw of the firm's objects, read against both groups, the firm's patents sit further beyond the randomization benchmark from other industries' patents than from its own industry's: the difference in excesses is 0.0051, outside the band from −0.0012 to 0.0012 that the joint draws give. The two

excesses are 0.0089 and 0.0037. This comparison rests on the distance between the firm's patents and another group's patents, whose side no ordering of Appendix A.15 sets, so it is a description.

Table G3: All other owners and the firms of other industries, against the randomization benchmark

		Event weights			Equal weights		
Distance	N	Excess	Share above	Holm p	Excess	Share above	Holm p
A. Every other owner: every patent; every linked repository once							
The firm's code and the others' code	2,472	−0.1215	0.19	0.004	−0.1213	0.159	0.004
The firm's code and the others' patents	2,472	0.0195	0.747	0.004	0.0271	0.791	0.004
The firm's patents and the others' code	2,472	0.1006	0.975	0.004	0.0988	0.978	0.004
The firm's patents and the others' patents	2,472	0.0082	0.708	0.002		(the same)	
B. Every firm the crosswalk matches; each firm's linked code summed							
The firm's code and the others' code	2,472	−0.1215	0.19	0.004	−0.1218	0.159	0.004
The firm's code and the others' patents	2,472	0.0193	0.747	0.004	0.0269	0.79	0.004
The firm's patents and the others' code	2,472	0.1006	0.975	0.004	0.0988	0.979	0.004
The firm's patents and the others' patents	2,472	0.0085	0.707	0.002		(the same)	
C. The firms of other industries							
The firm's patents and their patents	2,450	0.0089	0.718	0.002		(the same)	

The firm's objects are reassigned within its start-year strata against the other portfolios held as observed, with the firm left out of them; the test pools the firm-years, Holm-adjusted across the two weightings. Patents carry no repository weights, so the patent-to-patent rows are the same under both weightings.

F.4 The readings against the firm's own objects

Table G4 reads the four distances from the firm to its industry against the randomization benchmark, with the industry's portfolios held as observed. The firm's code sits nearer its industry's code than the randomization benchmark places it, and further from its industry's patents; its patents sit further from its industry's code, and further from its industry's patents. Where firm-clustered 95 percent intervals for the mean excess exist, they lie on the same side as the share test: for the firm's code against its industry's code, $[-0.1186, -0.0972]$ under event weights and $[-0.1241, -0.1025]$ under equal weights; for its code against its industry's patents, $[0.0552, 0.0684]$ and $[0.0725, 0.089]$.

Table G4: The four distances from the firm to its industry, against the randomization benchmark

		Event weights			Equal weights		
Distance	N	Excess	Share above	Holm p	Excess	Share above	Holm p
The firm's code and the industry's code	2,441	−0.1079	0.208	0.004	−0.1133	0.163	0.004
The firm's code and the industry's patents	2,448	0.0618	0.828	0.004	0.0808	0.89	0.004
The firm's patents and the industry's code	2,441	0.0949	0.973	0.004	0.0981	0.975	0.004
The firm's patents and the industry's patents	2,448	0.0037	0.57	0.002		(the same)	

The firm's objects are reassigned within its start-year strata; the industry's portfolios, with the firm left out, are held as observed. The test pools the firm-years, Holm-adjusted across the two weightings. Patents carry no repository weights, so the last row is the same under both weightings.

What depends on the firm's own choice of channels in these readings is bounded. By Lemma 15, the part of each reading that depends on which of its objects the firm made code is at most the channel's distance from the randomization benchmark's center: without start-year strata, the other channel's share times the firm's own distance, so it vanishes without the firm's own separation; with them, that bound moves by at most a shift the records do not compute. The rest of each

reading does not depend on that assignment. Two orderings of one distance beside another set the side: the firm's code sits further from its industry's patents than its patents do, and its patents sit further from its industry's code than its code does. Without strata, the lemma bounds those two readings below by them, each net of a sampling term that is never negative, and bounds above the excess of the distance between the firm's code and its industry's code. A check per firm-year, on equal weights, describes how the excesses sit beside the bounds; it is a description, not a test. For the firm's code against its industry's patents, excess and floor agree in sign in 80.8% of firm-years, against 70.4% for two independent signs with the same margins, and the excess is positive in 95.8% of the firm-years where the floor is positive (88.9% of all firm-years). For its patents against its industry's code they agree in 94.5% (92.8% if independent), and for its code against its industry's code in 84.2% (80.2%). The excess exceeds the bound without strata by more than three Monte Carlo standard errors in 0.6% of firm-years for one reading of the firm's code and in none for the other three. Beyond Monte Carlo error, only the strata can produce such an excess. The check was run on equal weights; the event-weight values of Table G4 stand beside it, not inside it.

The comparisons with all other owners fall under the same bound, at most twice the channel's distance from the randomization benchmark's center for a difference of two readings. Under one draw of the firm's objects, read against both its industry and all other owners, the difference in excesses, all other owners minus the industry, is -0.0423 for the firm's code against the patents and -0.0139 for its code against the code, under event weights; each lies outside the band its joint draws give. On equal weights these differences agree in sign with the difference of the floors in 83.7% and 83.2% of firm-years (60.7% and 50.1% if independent).

The reading of the firm's patents against its industry's patents falls under the bound too, but no ordering sets its side: excess and bound agree in sign in 59.5% of firm-years, against 46.2% if independent. It is a description, and so is the comparison with the firms of other industries that rests on it (Appendix F.3).

F.5 The firm and its industry against the within-industry benchmark

The within-industry benchmark is the declared reference for the firm's own distance and the industry's. The four distances from the firm to its industry are read against it here as descriptions (Table G5): no prediction is read against them. The firm's code against its industry's code depends on the weighting. Under event weights the firm's code sits nearer its industry's code than draws with the channels mixed place it, and under equal weights further, with smallest Holm-adjusted p of 0.04 and 0.04. The other three distances sit further apart than the within-industry benchmark places them, under both weightings.

F.6 Firms with and without linked code

The bound for the links the crosswalk misses runs on all 4,048 firms without linked code, by patent-count tercile, cut at 5 and 24 patents (section F.7). A public search under the linkage's own rule finds code the crosswalk misses for some firms without linked code, at rates of 39.7%, 32.6%

Table G5: The four distances from the firm to its industry, against the within-industry benchmark

		Event weights			Equal weights		
Distance	N	Excess	Share above	Holm p	Excess	Share above	Holm p
The firm's code and the industry's code	2,441	−0.0505	0.381	0.04	0.036	0.58	0.04
The firm's code and the industry's patents	2,448	0.1898	0.98	0.04	0.3004	0.996	0.04
The firm's patents and the industry's code	2,441	0.2734	0.967	0.04	0.3376	0.98	0.04
The firm's patents and the industry's patents	2,448	0.3074	0.993	0.02		(the same)	

The industry-year's objects, the firm's included, are reassigned within start years across channels and owners; the firm's portfolios and the industry's are read from each draw's slots. Share above and Holm p as in Table G1. Descriptions: no prediction is read against these rows.

and 61.4% by patent-count tercile, upper bounds that count name matches, not activity. Dividing each tercile's contrast by one minus its rate bounds the contrast with firms that hold no code, under assumptions that are not tested. On the patents' distance to the industry's code, observed and against the industry-pool benchmark, every tercile shares the sign, so the bound keeps the contrast on its side of zero. On the patents' distance to the industry's patents, the industry-pool benchmark's contrast is small in the lower two terciles and positive in the top one (−0.011, −0.007 and 0.042); the terciles disagree in sign, so no bound on its sign follows, and the reversal sits among the largest patent holders.

The comparison of firms with and without linked code uses every Compustat North America firm with a patent granted 2015–2024 in the patent crosswalk, in fiscal years 2020–2024: 18,201 firm-years. Of these, 17,755 firm-years of 4,087 firms, in 661 industry-years, have an industry group in the year before; the other 446 do not. Under the primary definition, a firm has linked code when the crosswalk links it to repositories that carry code mass: 3,561 firm-years of 787 firms, against 14,194 firm-years without. The definition is by firm, so it mixes years before and after a firm's code appears. Three companions sit beside it: the first counts only the firm-years with code in the window; the second drops the 253 firm-years of firms with linked code but none in the window; the third adds the firms with an accepted link and no code mass, giving 3,655 firm-years with linked code.

Each contrast regresses a distance, or its excess over a reference, on an indicator for linked code within industry-years, with standard errors clustered by firm. The industry-pool benchmark applies to every firm, since it redraws every firm's objects from its industry's pools while holding its counts. The randomization benchmark applies only where the firm has code in the window: a firm-year without code has nothing to reassign, and its excess is zero by construction. "No linked code" is the crosswalk's statement, not a finding about who publishes: the missed-link rates and the bound they give are in section F.7, and the industry's code pools miss what the crosswalk misses. Table G6 gives the within-industry-year differences under the four definitions, observed and in excesses over the industry-pool benchmark; their signs agree across definitions.

For the firms with code in the window, the randomization benchmark gives a reference of their own (Table G7). Their patents sit further from all other owners' code than the randomization benchmark of their own objects places them: the mean excess is 0.0831 under event weights (standard error 0.0036), on 3,114 firm-years of 772 firms. The table gives beside it their excesses over their

Table G6: Firms with and without linked code under four definitions: differences within industry-years

Distance	Primary	Code in the window	Without code-less years	With accepted links
A. Observed: with linked code minus without				
The firm's patents and the industry's code	−0.068 (0.005)	−0.071 (0.005)	−0.072 (0.005)	−0.067 (0.005)
The firm's patents and the industry's patents	−0.109 (0.008)	−0.113 (0.008)	−0.114 (0.008)	−0.107 (0.008)
The firm's patents and all other owners' code	−0.066 (0.005)	−0.071 (0.005)	−0.071 (0.005)	−0.066 (0.005)
The firm's patents and all other owners' patents	−0.067 (0.004)	−0.071 (0.004)	−0.071 (0.004)	−0.067 (0.004)
B. The difference in excesses over the industry-pool benchmark				
The firm's patents and the industry's code, event	−0.024 (0.004)	−0.024 (0.005)	−0.024 (0.005)	−0.023 (0.004)
The firm's patents and the industry's code, equal	−0.025 (0.005)	−0.025 (0.005)	−0.026 (0.005)	−0.025 (0.004)
The firm's patents and the industry's patents	0.087 (0.008)	0.091 (0.008)	0.091 (0.008)	0.087 (0.008)

Differences within the industry-year, firms with linked code minus firms without, with standard errors clustered by firm in parentheses; the industry's portfolios leave the firm out, and all other owners' portfolios are every other owner's. Panel A is observed under event weights; panel B is the same regression with each firm's excess over the industry-pool benchmark in place of its distance. A positive difference in excesses says the firms with linked code sit further apart, relative to the industry-pool benchmark, than the firms without.

own industry's portfolios, under the same randomization benchmark and under the industry-pool benchmark. These are descriptions of the firms with code, not comparisons with the firms without.

Table G7: Firms with code in the window: the patents' excesses over three references

Distance	Reference	N	Firms	Mean excess	Firm's share
The patents and all other owners' code, event	Own objects	3,114	772	0.0831 (0.0036)	0.0013
The patents and all other owners' code, equal	Own objects	3,114	772	0.0796 (0.0034)	0.0011
The patents and all other owners' patents	Own objects	3,114	772	0.0053 (0.0006)	0.0007
The patents and listed owners' code, event	Own objects	3,114	772	0.0831 (0.0036)	0.0013
The patents and listed owners' patents	Own objects	3,114	772	0.0055 (0.0006)	0.0007
The patents and the industry's code, event	Own objects	3,100	771	0.0787 (0.0034)	
The patents and the industry's code, equal	Own objects	3,100	771	0.0806 (0.0034)	
The patents and the industry's patents	Own objects	3,112	772	0.0012 (0.0016)	
The patents and the industry's code, event	Industry pool	3,100	771	0.0532 (0.0044)	
The patents and the industry's patents	Industry pool	3,112	772	0.2529 (0.0065)	

Firm-years of firms with linked code that have code in the window. An excess is the observed distance minus the reference's mean; standard errors clustered by firm in parentheses. "Own objects" is the randomization benchmark of the firm's objects within its start-year strata against the other portfolios held as observed, with the firm left out; "Industry pool" is the industry-pool benchmark. The firm's share is its mean share of the pool that includes it. Listed owners are every firm the crosswalk matches.

F.7 Firms with and without linked code

Do the patents of firms with linked code sit differently within their industry from those of firms without? Against the industry-pool benchmark, within industry-years, firms with linked code hold patents less separated from their industry's code than firms without, and more separated from its patents. The year a firm starts to publish code goes with no excess distinguishable from zero. A bound for the links the crosswalk misses keeps the sign of the first contrast only. The comparison uses the two distances that exist for any patenting firm, over the universe of Table C5. A firm with

linked code counts as such in every year, before and after its code appears.

Raw, the patents of firms with linked code sit nearer their industry's patents, as they sit nearer its code; against the industry-pool benchmark the second comparison turns round. Firms with linked code hold more patents, and the more objects a portfolio holds, the nearer it lies to any pool under random assignment alone, the small-unit bias (D'Haultfœuille and Rathelot, 2017, p. 40). The industry-pool benchmark holds each firm's counts fixed, and with them fixed the patents of firms with linked code sit further from their industry's patents (Table G8, panel A). Each of the two contrasts against the industry-pool benchmark has a Holm-adjusted p of < 0.001, in a family of three. The third member asks whether the year a firm starts to publish code goes with its excess, with the excess as the dependent variable: among the firms at risk in spell years 2021–2024, those that start in a year hold patents slightly nearer their industry's code, against the industry-pool benchmark, than at-risk firms that do not, within the industry-year and given the hazard's covariates. This is the direction of the first contrast, but it is not distinguishable from zero (Holm-adjusted p 0.164, panel B).

Limits. The reversal holds among the largest patent holders, and under a bound for the links the crosswalk misses its sign can no longer be pinned down within patent-count terciles, while the sign of the contrast with the industry's code is kept (Appendix F.6). "No linked code" is the crosswalk's statement, not a finding about who publishes, and the industry's code pools also miss the code the crosswalk misses; each contrast is an association of crosswalk status with position within industry-years, not an effect of opening.

Table G8: Firms with and without linked code, within industry-years

A. Firms with linked code minus firms without, within the industry-year

Distance	Firm-years (with code)	Cells	Observed difference	Difference in excesses over the industry-pool benchmark	Holm p
The firm's patents and the industry's code, event weights	15,407 (3,313)	456	−0.068 (0.005)	−0.024 (0.004)	< 0.001
The firm's patents and the industry's code, equal weights	15,407 (3,313)	456	−0.079 (0.005)	−0.025 (0.005)	—
The firm's patents and the industry's patents	15,735 (3,325)	467	−0.109 (0.008)	0.087 (0.008)	< 0.001

B. The year a firm starts to publish code: the excess on the left

	Firm-years (firms)	Starts	Coefficient on starting, the excess over the industry-pool benchmark	Holm p
The firm's patents and the industry's code, event weights	8,387 (2,636)	72	−0.013 (0.009)	0.164

Panel A: the universe of patenting firms with an industry group (Table C5), at the firm-years where the row's distance is defined, so each row holds fewer firm-years than the universe and the two rows differ; each row regresses the distance, or its excess over the industry-pool benchmark, on an indicator for linked code, absorbing the industry-year, with standard errors clustered by firm in parentheses; cells are the industry-years holding both kinds of firm. An excess is the observed distance minus the reference's mean. Panel B: the entry hazard's firms at risk in spell years 2021–2024; the excess is regressed on an indicator for starting to publish code in the year, absorbing the industry-year, with the hazard's covariates. Holm's adjustment runs over a family of three: the two event-weight contrasts of panel A against the industry-pool benchmark and panel B's coefficient. Appendix Table G6 gives the contrasts under three other definitions of linked code.

G Nearest neighbors

Does counting a firm's open-source code change which firms are its nearest neighbors? It changes more of them than the randomization benchmark does. A firm's technological position is how its patents spread over the main groups, and its nearest neighbors, which I call here its *nearest firms*, are the firms with the highest technological proximity to it, the cosine of two positions (Jaffe, 1986). Studies of knowledge spillovers and of technological rivalry pair firms in this way (Jaffe, 1986; Bloom et al., 2013). On the comparison sample's 595 firms and 2,472 firm-years, I recompute each firm's position with its open-source code counted, each item once, and count how many of its ten nearest firms change (Table G9). The randomization benchmark gives the same count after dealing out again, for every firm at once, the technologies of each firm's own patents and open-source code at random.

Under event weights, 5.67 of a firm's ten nearest firms change once its open-source code counts, against 4.68 under the randomization benchmark. The excess is 0.99, with a 95 percent interval from 0.89 to 1.1; under equal weights it is 1.7. The whole map of proximities reorders too. When I measure the proximity of every pair of firms twice, from patents alone and with open-source code counted, the rank correlation between the two is 0.669, against 0.888 under the randomization benchmark. The excess is of similar size across the terciles of the count of open-source projects (1.04, 0.77 and 1.17) and across sectors (from 0.62 to 1.3), so it is not confined to the largest programs. Counting its open-source code moves a firm's own position 2.04 times as far as the randomization benchmark moves it (panel C). This move is the *shift*: one minus the proximity between the firm's position from patents alone and its position with its open-source code counted. The shift beyond the randomization benchmark is the excess of section 4.2, seen from the patent side.

When one language model labels both channels, on a sample of 45 firms, 0.807 of the shift's excess remains under event weights (interval 0.606 to 1.008) and 0.858 under equal weights (0.72 to 0.996). Both intervals span 0.8, so the share lies between the two bands, the one where most of the excess remains and the one where part of it comes from the labels. The event-weight point estimate falls just below 0.8, and the equal-weight one above it. The intervals leave room for part of it to come from the two channels' different classifications (panel D).

Work on spillovers and rivalry would pair firms differently once open-source code counts. That work pairs firms by patent proximity, but a firm that publishes open-source code does not keep all of its nearest firms once its open-source code counts: about one nearest firm in ten changes beyond the randomization benchmark.

Table G9: Counting a firm's OSS: which firms sit nearest to it, and how far its position moves

Statistic	Observed	Benchmark [band]	Excess [interval]
A. The firm's nearest firms: technological proximity from patents alone and with OSS counted			
Of the ten nearest firms, how many change, event weights	5.67	4.68 [4.56, 4.78]	0.99 [0.89, 1.1]
Of the ten nearest firms, how many change, equal weights	5.25	3.55 [3.47, 3.64]	1.7 [1.59, 1.8]
Share of firm-years whose nearest firm changes, event weights	0.657	0.579 [0.561, 0.595]	0.078 [0.061, 0.095]
Rank correlation of the proximities between firms, event weights	0.669	0.888 [0.877, 0.9]	
Rank correlation of the proximities between firms, equal weights	0.66	0.912 [0.904, 0.919]	
pairs in the same industry, event weights	0.73	0.874 [0.86, 0.886]	
other pairs, event weights	0.632	0.889 [0.874, 0.901]	
B. Of the ten nearest firms, how many change: by the count of OSS projects and by sector (event weights)			
OSS projects active in the window: at most 19	5.21	4.17	1.04 [0.89, 1.19]
more than 19, at most 80	5.43	4.66	0.77 [0.6, 0.94]
more than 80	6.36	5.2	1.17 [0.97, 1.37]
Information (NAICS 51)	6.39	5.52	0.88 [0.69, 1.06]
Manufacturing (NAICS 33)	4.14	3.15	1.0 [0.82, 1.17]
Manufacturing (NAICS 32)	5.05	3.99	1.05 [0.69, 1.41]
Professional, scientific and technical services (NAICS 54)	7.17	5.92	1.25 [0.9, 1.59]
Finance and insurance (NAICS 52)	5.45	4.32	1.13 [0.8, 1.46]
Retail trade (NAICS 45)	5.9	4.9	1.0 [0.12, 1.88]
Other sectors	7.18	5.88	1.3 [0.99, 1.61]
No NAICS code	3.98	3.36	0.62 [0.2, 1.05]
C. The same excess seen from the patent side: how far counting OSS moves the firm's own position			
Event weights	0.2293	0.1126 [0.1078, 0.1172]	0.1167 [0.1039, 0.1295]
Equal weights	0.1953	0.0582 [0.0559, 0.0607]	0.1371 [0.123, 0.1512]
Observed shift over the randomization benchmark's mean	2.04 under event weights; 3.36 under equal weights		
The patent drift in proximity, 2020–24 against 2015–19	0.1868	0.0984	0.0884 [0.0764, 0.1004]
The shift's excess minus the drift's, event weights			0.035 [0.0141, 0.0559]
The shift's excess minus the drift's, equal weights			0.0516 [0.0302, 0.0729]
D. One language model on both channels: the share of the shift's excess that remains			
Event weights	0.807		[0.606, 1.008]
Equal weights	0.858		[0.72, 0.996]

OSS means open-source software. Panels A to C: the comparison sample, 595 firms and 2,472 firm-years, five-year windows ending in 2020–2024. P_{it} is the firm's technological position from its patents alone (each patent's weight split equally over its main groups). O_{it} is its open-source code's position (event or equal weights). With the open-source code counted, every item counts once: $C_{it} = (1 - \lambda_{it})P_{it} + \lambda_{it}O_{it}$. Here λ_{it} is the open-source code's share of the window's items (mean 0.385). J is technological proximity (Jaffe, 1986), the cosine of two positions over the main groups. A firm's ten nearest firms are the ten other firms of the sample in the same year with the highest J. A count is ten minus the number of firms in both lists. For every firm at once, the randomization benchmark deals out again the technologies of each firm's own patents and open-source projects among them within start years. The randomization benchmark column gives the mean over the draws. In brackets, it gives the band holding the central 95 percent of the draws' means (pooled panels only). The excess is the observed statistic minus the randomization benchmark's mean, with its firm-clustered 95 percent interval. Panel C restates the paper's excess from the patent side, since $D(P_{it}, C_{it}) = \lambda_{it} D_{it}$ exactly. The patent drift in proximity compares each firm's patents of two windows that do not overlap. It uses the 353 firms that have their 2024 window in the sample and are in the panel in 2019. Its reference is a randomization benchmark that redraws those patents across the two windows (one row per firm, robust standard errors). Panel C's shift is one minus J between the firm's position from patents alone and its position with OSS counted. The drift row is one minus J between the firm's patents of 2015–2019 and of 2020–2024. The main text measures the same patent drift by the Bar–Leiponen distance. D_{it} is the Bar–Leiponen distance between the firm's open-source code and its patents. Panel D: one language model labels both channels of a 45-firm sample (185 firm-years). The ratio's numerator is the shift's excess with the model's classes on both channels. Its denominator is the excess with the examiners' classes on the patents and the labels on the open-source code. The ratio has a design-based interval.

H The six distances in full

This appendix gives the six distances of section 5.5 with their benchmark distances, the regressions on them in full, and the earlier work they bear on.

H.1 The six distances and their benchmark distances

A firm's open-source code sits nearer its industry's open-source code than its industry's patents, yet nearer still under the industry-pool benchmark (Table G10). The mean distances are 0.6199 to the industry's open-source code and 0.8576 to its patents, and the industry-pool benchmark gives 0.5461 for the first. That benchmark deals the firm's open-source projects again from its industry's open-source code. Since the observed mean distance lies above the benchmark distance, a firm's open-source code keeps a mix of its own within its industry's open-source code, which the industry's combined portfolios do not show. One distance, between the firm's patents and its industry's patents, is the same under both weightings, so its family of tests in section 5.5 has five tests. Appendix E.9 uses the industry-pool benchmark to ask whether a firm's patents sit differently in its industry in the year it starts to publish open-source code.

Table G10: The six distances

Distance		Mean	SD	Median	Firm-years
The firm's patents and its OSS	Observed	0.8062	0.1641	0.8444	2,450
	The industry-pool benchmark	0.8293	0.1212	0.8425	
	The within-industry benchmark	0.7018	0.159	0.7219	
The firm's OSS and the industry's OSS	Observed	0.6199	0.1859	0.6063	2,441
	The industry-pool benchmark	0.5461	0.1472	0.5443	
The firm's OSS and the industry's patents	Observed	0.8576	0.1077	0.872	2,448
	The industry-pool benchmark	0.8139	0.1285	0.8267	
The firm's patents and the industry's OSS	Observed	0.81	0.1377	0.8304	2,441
	The industry-pool benchmark	0.7462	0.1497	0.7518	
The firm's patents and the industry's patents	Observed	0.6784	0.1821	0.6892	2,448
	The industry-pool benchmark	0.3794	0.2077	0.3551	
The industry's patents and its OSS	Observed	0.7247	0.1605	0.6968	2,439
	The industry-pool benchmark	0.6982	0.1731	0.7029	
	The within-industry benchmark	0.4179	0.187	0.3837	

OSS means open-source software. SD is the standard deviation. Event weights; the comparison sample's firm-years with an industry. The industry's portfolios leave the firm out. For each firm-year, a benchmark's row takes the benchmark distance: the distance's mean over that benchmark's draws. The row then gives the mean, SD and median of those firm-year values. The industry-pool benchmark deals the technologies of the industry-year's items again within each channel, the firm's own included. Each portfolio then becomes a draw of its own size from its channel's pool. The within-industry benchmark deals the same technologies again across channels and owners. Section 5.5 compares four of these distances with their benchmark distances. Section 5.5 asks what goes with the excess of each distance that involves the industry. Appendix F.1 gives equal weights.

Without a randomization benchmark, earlier work also finds that firms keep a mix of their own within their industry. Firms in the same standard industry often offer different products

(Hoberg and Phillips, 2016, p. 1435), and firms competing with similar products often rely on different technologies (Arts et al., 2023, p. 7255). Stuart and Podolny (1996, p. 31) find that ten Japanese semiconductor firms hold distinct technological positions within one industry. Patel and Pavitt (1995, p. 20) qualify this: across 34 broad fields, a large firm is more likely to find a similar patent profile inside its sector than outside it. None of these compares a firm with its industry's portfolios against a randomization benchmark, and none measures it for open-source code. Because a firm reveals a selected part of its software as open-source code (Henkel, 2006, p. 961), I read its open-source code here as what it opens, not as its whole technology.

H.2 The regressions on each distance

Section 5.5 runs the regression of equation (1) on each of the five distances that involve the industry, with only the firm's three counts and the fixed effects on the right:

$$Y_{it}^{k} = \beta_R^k \log(1+n_{i,t-1}^{\text{oss}}) + \beta_E^k \log\Big(1+\frac{n_{i,t-1}^{\text{ev}}}{n_{i,t-1}^{\text{oss}}}\Big) + \beta_P^k \log(1+n_{i,t-1}^{\text{pat}}) + \phi_{it} + \varepsilon_{it}, \qquad Y_{it}^{k} \in \{D_{it}^{k},\ \mu_{it}^{k},\ \tilde{D}_{it}^{k}\}. \tag{7}$$

The index k names one of the five distances. Four of them run from one of the firm's portfolios to one of its industry's: two from the firm's open-source code, to its industry's open-source code and to its industry's patents, and two from the firm's patents, to the same two portfolios. The fifth is the industry's own distance, between its patents and its open-source code. D_{it}^{k} is the observed distance, and μ_{it}^{k} is its benchmark distance, the mean over its benchmark's 499 draws. The four distances from the firm's portfolios use the industry-pool benchmark, and the industry's own distance uses the within-industry benchmark (section 5.5). $\tilde{D}_{it}^{k} = D_{it}^{k} - \mu_{it}^{k}$ is the excess. The counts are those of equation (1), and β_R^k, β_E^k and β_P^k are their slopes for distance k. ϕ_{it} is either firm and year effects, with which a slope compares each firm with itself over time, or industry-by-year effects, with which a slope compares firms within the same industry-year. ε_{it} is the error, and I cluster the standard errors by firm. The regressions use the 2,413 firm-years of 585 firms with an industry and all six distances that the fixed effects leave identified (section 3.6). They describe; none is a test declared before testing.

Within a firm and at the firm's own counts, each excess rises with the count of the channel its distance runs from, because the benchmark distance falls faster with that count than the observed distance does (Table G12). Each benchmark draws as many items as the firm holds in the channel the distance runs from, so that count moves the benchmark distance most. A firm's open-source code, for example, comes nearer its industry's open-source code as the firm publishes more of it: the observed distance changes by about -0.025 (0.005) when the firm's open-source projects double. The industry-pool benchmark would bring the firm's open-source code nearer still, with a change of about -0.038 (0.004), so the excess rises. The firm's patents do the same against its industry's patents: the observed distance changes by about -0.023 (0.003) when the firm's patents double, against about -0.065 (0.003) for the benchmark distance, so that excess rises as well.

Across firms in the same industry-year and at the firm's own counts, the two excesses from

the firm's open-source code again rise with its open-source projects (Table G11). In a firm with twice the open-source projects of another firm in its industry-year, the excess of $D(O_{it}, O_{G,t})$, to its industry's open-source code, is higher by about 0.014 (0.003), and the excess of $D(O_{it}, P_{G,t})$, to its industry's patents, by about 0.010 (0.001). Both excesses from its patents rise with its patents, the excess of $D(P_{it}, P_{G,t})$ more. In a firm with twice another's patents, the excess of $D(P_{it}, P_{G,t})$, to its industry's patents, is higher by about 0.029 (0.002), and the excess of $D(P_{it}, O_{G,t})$, to its industry's open-source code, rises slightly, by about 0.004 (standard error 0.002, p 0.018). In a firm with twice another's open-source projects, both excesses from its patents are lower, by about -0.011 (0.003) and -0.011 (0.003). The placebo world of section 5.2.3 gives a positive sign to the slope of the excess of $D(P_{it}, O_{G,t})$ on the patents, and to the two slopes on the open-source projects above, at the firm's counts. The data's slopes lie below that world's. Across firms, then, a larger open-source program goes with larger excesses from the firm's open-source code and smaller ones from its patents, each at the firm's own counts. At equal size (section 5.5), none of the four excesses rises with the count of open-source projects: across firms, the two from the firm's open-source code change by -0.003 (standard error 0.002) and -0.00043 (0.001) when its open-source projects double.

Table G11: The distances between the firm and its industry on the firm's OSS counts, across firms in an industry-year

	Firm's OSS to industry's OSS	Firm's OSS to industry's patents	Firm's patents to industry's OSS	Firm's patents to industry's patents	Industry's patents to industry's OSS
A. The observed distance					
Log OSS projects (lagged)	−0.059***	−0.031***	−0.023***	−0.012***	−0.005***
	(0.005)	(0.002)	(0.004)	(0.004)	(0.002)
Log events per OSS project (lagged)	0.006**	0.006***	−0.007	−0.002	−0.00022
	(0.005)	(0.002)	(0.004)	(0.004)	(0.002)
Log patents (lagged)	−0.002	0.002**	−0.012***	−0.056***	0.003***
	(0.003)	(0.001)	(0.002)	(0.003)	(0.001)
B. The benchmark distance (mean over the benchmark's 499 draws)					
Log OSS projects (lagged)	−0.078***	−0.046***	−0.007***	0.005*	−0.008***
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)
Log events per OSS project (lagged)	0.023***	0.011***	−0.002	−0.005	−0.002
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Log patents (lagged)	0.001	0.001	−0.017***	−0.097***	0.006***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
C. The excess over the benchmark (the observed distance minus the benchmark distance)					
Log OSS projects (lagged)	**0.020****	**0.015****	−0.015***	−0.016***	0.003**
	(0.005)	(0.002)	(0.004)	(0.004)	(0.002)
Log events per OSS project (lagged)	−0.016***	−0.005***	−0.005	0.003	0.001
	(0.005)	(0.002)	(0.004)	(0.005)	(0.001)
Log patents (lagged)	−0.004	0.001*	**0.005***	**0.042****	−0.003**
	(0.003)	(0.001)	(0.002)	(0.003)	(0.001)
Firm-years	2,413	2,413	2,413	2,413	2,413
Firms	585	585	585	585	585

OSS means open-source software. Under event weights, each column regresses one distance between the firm and its industry, its benchmark distance and its excess on the firm's lagged counts. Each count is over the five-year window that ends the year before the distance's window ends. Each count enters as the log of one plus the count, so a coefficient is the change per unit of that log. The text reads a coefficient times $\ln 2$, the change when the count about doubles. The industry's portfolios leave the firm out. The four distances from the firm's portfolios use the industry-pool benchmark. The industry's own distance uses the within-industry benchmark (section 5.5). The benchmark distance is the mean distance over the benchmark's draws. The excess is the observed distance minus the benchmark distance. In every column, panel C's coefficient equals panel A's minus panel B's. Bold marks the count of the channel the distance runs from. The industry's own distance runs from neither. The sample is every firm-year with an industry and all six distances that the fixed effects leave identified. With industry-by-year effects, each slope compares firms within the same industry-year. Parentheses show standard errors clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from standard errors clustered by firm.

Earlier work supports the reason for the rise, not the rise itself. Carrington and Troske (1997, p. 405) and Allen et al. (2015, p. 45) show that a dissimilarity index is inflated when its units are small and that the inflation shrinks as the units grow. For patent portfolios, Benner and Waldfogel (2008, NBER Working Paper 13322, p. 17) show that distances computed from few patents are biased upward when firms are truly close, and Gentzkow et al. (2019b, p. 1320) find the same for differences between groups' choices. That is why each benchmark distance draws the firm's own number of items, and why the count of the distance's own channel moves the benchmark distance most. D'Haultfœuille and Rathelot (2017, p. 62) qualify this: an index read against random

allocation can still vary with unit size when the underlying separation does not. The excess rises because the benchmark distance falls with the counts and the observed distance falls less; under the industry-pool benchmark, the two would fall together.

Earlier work also finds that firms keep their technology mixes. Patel and Pavitt (1995, p. 16) find large firms' patent profiles across 34 broad fields remarkably stable between 1969–74 and 1985–90, and Stuart and Podolny (1996, p. 28) find that the technological positions of ten Japanese semiconductor firms changed little while the firms grew. Arts et al. (2023, p. 7258) find that differences between firms account for most of the variance in technology differentiation and that changes within a firm account for little of it. Teece et al. (1994, p. 9) and Breschi et al. (2003, p. 82) find that firms stay coherent, against a random reference, as they spread over more activities. In levels, larger firms sit closer to other firms in technology space (Bloom et al., 2013, p. 1385), and larger, older firms are also less differentiated from their neighbors (Arts et al., 2023, p. 7254). Jaffe (1989, p. 90) limits the persistence: about a third of firms changed technology cluster within about a decade. Against a randomization benchmark, none of these works reads how a firm's portfolios come nearer its industry's, so none can say whether they come nearer more slowly than the counts alone would bring them. The excess says that they do. Table G12 gives the same regressions within a firm.

Table G12: The distances between the firm and its industry, and the firm's counts, within a firm

	From the firm's open-source code to the industry's		From the firm's patents to the industry's		Within the industry
	Open-source code	Patents	Open-source code	Patents	Patents to open-source code
	(1)	(2)	(3)	(4)	(5)
Panel A. Observed distance					
Log open-source projects	−0.037***	−0.016***	−0.003	−0.005	−0.002
	(0.007)	(0.003)	(0.003)	(0.004)	(0.002)
Log events per project	0.016***	0.006***	−0.005	−0.005	−0.003**
	(0.006)	(0.002)	(0.002)	(0.003)	(0.001)
Log patents	0.007	−0.004	−0.012**	−0.033***	−0.002
	(0.007)	(0.003)	(0.004)	(0.005)	(0.002)
Panel B. Benchmark distance (mean over the benchmark's 499 draws)					
Log open-source projects	−0.054***	−0.028***	−0.009**	−0.005	−0.010**
	(0.006)	(0.002)	(0.002)	(0.004)	(0.004)
Log events per project	0.018***	0.009***	0.00014	−0.001	−0.002
	(0.004)	(0.002)	(0.001)	(0.003)	(0.002)
Log patents	0.008*	−0.003*	−0.027***	−0.094***	0.004
	(0.004)	(0.002)	(0.002)	(0.004)	(0.003)
Panel C. Excess over the benchmark (observed minus benchmark distance)					
Log open-source projects	**0.018****	**0.012****	0.006	0.0000032	0.008**
	(0.007)	(0.003)	(0.003)	(0.004)	(0.004)
Log events per project	−0.002	−0.003***	−0.005	−0.004	−0.001
	(0.005)	(0.002)	(0.003)	(0.003)	(0.002)
Log patents	−0.001	−0.00047	**0.015****	**0.060****	−0.006
	(0.007)	(0.003)	(0.004)	(0.004)	(0.003)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,413	2,413	2,413	2,413	2,413
Firms	585	585	585	585	585

Notes: The table reports estimates of equation (1), with only the three counts and firm and year fixed effects as regressors and with the distances under event weights. Each column uses one distance that involves the industry. The outcome is the observed distance in Panel A, the benchmark distance (the mean distance over the benchmark's draws) in Panel B, and the excess (the observed distance minus the benchmark distance) in Panel C. Each coefficient in Panel C therefore equals the one in Panel A minus the one in Panel B. Each count covers the five-year window that ends the year before the outcome's window ends and enters as the log of one plus the count. The text multiplies a coefficient by $\ln 2$, approximately the change when the count doubles. The industry's portfolios exclude the firm. The four distances from the firm's portfolios use the industry-pool benchmark, and the industry's own distance uses the within-industry benchmark (section 5.5). Bold marks the count of the channel from which the distance is measured. The industry's own distance is measured from neither of the firm's channels. The sample is every firm-year with an industry and all six distances that the fixed effects leave identified. With firm and year fixed effects, each coefficient compares a firm with itself over time. In column 1, the observed distance from the firm's open-source code to its industry's open-source code falls with the firm's count of open-source projects, as the small-unit bias makes it fall (section 5.1), but more slowly than its benchmark distance. Standard errors clustered by firm are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided, from standard errors clustered by firm.

I The analysis plan

Each test the paper reads was fixed in a dated record before its statistic was computed: the statistic, its reference, the family over which Holm's adjustment runs and the rule that turns the statistic into a reading. Table H1 lists the records, each at its path in the project's repository. A record changes only through a dated amendment, kept beside it in the same folder, and the results are reported under each record and its amendments as they stand.

Table H1: The records that fixed each test before its statistic was computed

What was fixed	Where it is read	Record
The randomization benchmark, its share test and its Holm family	Section 4.3	`topics/T02-distance-and-chance-benchmarks/code/contracts/object_shuffle_specification.json` (28 September 2026)
The worlds without behavior, the fixed-portfolio world among them	Section 4.4	`topics/T02-distance-and-chance-benchmarks/code/contracts/no_behaviour_worlds_specification.json` (27 September 2026)
The outside-inside split of the excess and its four readings	Section 4.3; Appendix A	`topics/T02-distance-and-chance-benchmarks/code/contracts/gap_split_specification_20261004.json` (4 October 2026)
The technology-position world	Section 4.4	`topics/T02-distance-and-chance-benchmarks/code/contracts/propS_world_specification_20261004.json` (4 October 2026)
Close rivals: the scores, the peer groups and their Holm family; the common-rate world	Section 4.4; Appendix E.8	`topics/T04-technology-allocation/code/contracts/technology_or_firm_specification.json` (1 October 2026)
The order of the channels: the cells, the clocks and the permutation	Appendix E.8	`topics/T06-timing-prior-art-and-knowledge-use/code/contracts/t06_m5_order_timing_20261003.json` (3 October 2026)
The regressions with a distance on the left and their primary family, since withdrawn by a dated amendment beside it	Appendices B.8 and B.9	`topics/T03-firm-level-separation/code/contracts/distance_left_specification_20261008.json` (8 October 2026)
The six distances, their references, the share test that governs every row and its Holm family	Appendix H; Appendix A.13	`topics/T02-distance-and-chance-benchmarks/code/contracts/six_edge_full_run_20261010.json` (10 October 2026)
The stated signs for the six distances and the rule that reads them, fixed before those comparisons were computed	Appendix E.8	`topics/T08-theory-of-channel-choice/code/contracts/four_gates_20261010.json` (10 October 2026)
The stated signs for the paper's own tests, each recorded from its source with page and quote, after those tests had run	Appendix E.8	`topics/T10-literature-and-replication/output/p1_reasons_sources_20261011/sources.json` (11 October 2026)
Firms with and without linked code: the contrasts and their family of three	Appendix F.7	`topics/T07-who-opens-and-what-follows/code/contracts/b2_declaration_20261010.json` (10 October 2026)
The masking check and its materiality rule	Appendix C	`topics/T01-labels-and-linkage/code/contracts/tp01_masking_specification.json` (2 October 2026)
The scoring of the labels against the EPO's categorizer	Section 3.3; Appendix C	`topics/T01-labels-and-linkage/code/contracts/tp01_epo800_scoring_20261005.json` (5 October 2026)

Each record but the sources' record states that it was declared before any output of its stage; the re-runs on the corrected event dates of section 3.1 are recorded in dated run sheets and amendments in the same folders.