Actionable Disclosure in Patents: How Open Code Reduces Spatial Barriers to Knowledge Diffusion

Sergio Petralia & Ron Boschma

Papers in Evolutionary Economic Geography

26.11

Utrecht University Human Geography and Planning

Actionable Disclosure in Patents: How Open Code Reduces Spatial Barriers to Knowledge Diffusion

Sergio Petralia *[1], Ron Boschma[1]

[1]Department of Economic Geography, SGPL, Utrecht University, Utrecht, Netherlands.

*Corresponding author. Email: s.g.petralia@uu.nl

July 9, 2026

Abstract

The patent system rests on a fundamental bargain: temporary monopoly rights in exchange for the disclosure of technical knowledge. Yet, the complexity of transferring technical know-how often limits the effectiveness of disclosure, keeping innovation highly localized. We study whether actionable disclosure in the form of publicly available executable code can overcome these frictions. We rely on a novel dataset linking two decades of patenting activity in the United States to contributions in public code repositories by 1,556 organizations. Using neural language models, we measure the semantic similarity between patent descriptions and public code contributions by these firms to identify patents with high digital disclosure. We find that these patents attract citations from inventors located approximately 17% farther away than those citing a group of control patents, suggesting that actionable disclosure in the form of executable code reduces spatial barriers to knowledge diffusion.

Keywords: patent disclosure, knowledge diffusion, open-source software, geography of innovation, spatial spillovers

JEL Codes: O31, O33, O34, R12

1 Introduction

Patents represent a valuable source of knowledge. Patent literature ranks second only to customers and users as a source of innovation, with importance comparable to that of scientific literature (Giuri et al., 2007; Sampat, 2018). Approximately 80% of companies consider patent information important (OECD, 2004), and in rapidly advancing fields such as nanotechnology, 64% of academic researchers report having read a patent (Ouellette, 2011). Effective disclosure of this information lies at the core of the so-called patent bargain, the principle that an inventor receives a limited monopoly in exchange for public disclosure. By making technical information publicly available, the patent system creates a formal channel through which knowledge can diffuse beyond where it was originally developed. Unlike knowledge transmitted through personal interactions and local networks, patent documents are, in principle, accessible to inventors regardless of their geographic proximity to the original invention. Such disclosure should enable others to build upon existing inventions wherever they are located, fostering long-term productivity and innovation (Machlup and Penrose, 1950).

A substantial number of empirical studies have documented the virtuous role of disclosure in shaping knowledge diffusion (de Rassenfosse, 2025). A significant share of this evidence draws on the American Inventors Protection Act of 1999 (AIPA), which mandated the publication of United States (US) patent applications 18 months after filing rather than only upon grant. Hegde and Luo (2018) use this policy change to show that, in the biomedical sector, AIPA reduced licensing delays by approximately 10 months on average. Using the same reform, Baruffaldi and Simeth (2020) find that earlier disclosure accelerates information diffusion, with a one-year reduction in publication lag leading to a 12% increase in patent citations. In addition, Lück et al. (2020) document a decline in duplicative inventive effort, measured by a reduction in blocked patent claims. Likewise, Hegde et al. (2023) show that post-AIPA patents receive more and faster citations and face less duplication. Conversely, when information is withheld, whether due to government-imposed secrecy orders (Gross, 2019; de Rassenfosse et al., 2024) or strategic patenting practices (Bessen, 2005), technological diffusion is restricted. Gross (2023) provides particularly compelling evidence of such costs, showing that compulsory secrecy during World War II had persistent negative effects on affected firms' patenting activity.

However, the benefits of improved access to information appear to decay rapidly with distance. For instance, Furman et al. (2021) and Berkes and Nencka (2024) show that the establishment of patent libraries leads to significant increases in local patenting activity, but that these effects are primarily concentrated in their immediate vicinity. Similarly, Baruffaldi and Simeth (2020) observe that early disclosure tends to amplify knowledge diffusion mainly within familiar technological domains and established networks.

Knowledge flows remain strongly localized, as inventors disproportionately build upon geographically proximate knowledge (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014; Ganguli et al., 2020). One explanation is that knowledge often spreads through informal interactions (Andrews, 2023; Singh, 2005), as well as through local networks and local labor mobility (Breschi and Lissoni, 2009; Klevorick et al., 1995; Almeida and Kogut, 1999), which allow inventors to obtain complementary knowledge that is not completely captured in formal documents. In other words, geographic proximity is particularly valuable when disclosed information is incomplete, ambiguous, or difficult to translate into practice.

This relationship between tacit knowledge and proximity highlights a fundamental distinction between making knowledge available and making it actionable (Murray and O'Mahony, 2007). Patent documents make inventions publicly available, but their legal language, technical jargon, and limited implementation details may prevent inventors from independently reproducing or building upon them (Seymore, 2009; Fromer, 2008; Büttner et al., 2022). Disclosure's ability to diffuse knowledge across space may therefore depend on whether its form enables distant inventors to understand, test, and apply the underlying knowledge without access to localized tacit knowledge. While existing research has mostly studied the effect of changes on whether or when disclosure occurs de Rassenfosse (2025), much less is known about how more actionable forms of disclosure can substitute for geographic proximity.

In this article, we ask whether making disclosed knowledge more actionable can reduce the geographic frictions that constrain knowledge diffusion. We take advantage of a unique feature of certain patents to estimate how the public availability of detailed technical information, beyond what is disclosed in the patent document, affects the transfer of knowledge over space. We leverage the fact that certain patented technologies depend on public digital infrastructure, which often requires firms to modify existing code and contribute those (or part of those) improvements back to the original repositories to maintain full functionality (Henkel, 2006; Petralia, 2025). In these cases, publicly accessible technical information is released along with the patent. Executable code represents an unusually actionable form of disclosure. Unlike patent prose, it provides a precise and inspectable representation of an invention's technical workings and allows inventors to test, modify, and build upon it directly.

To do so, we rely on a novel database that links two decades (2005-2024) of contributions to 98 of the most popular public code repositories with patent filings in the United States for 1556 organizations.[1] We compare the geographic distance between citing and cited inventors across two groups of cited patents: those accompanied by publicly available code, and a matched control group without such code. The matching procedure is based on the technological classification of the patent, the number of inventors, assignees, and the filing date. This approach is similar to that of Hegde et al. (2023), who assess the impact of disclosure by comparing patents subject to stricter disclosure regulations with 'twin' counterfactual patents. It also closely follows prior literature that uses matching procedures to measure the localization of spillovers (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014; Ganguli et al., 2020).

Measuring the extent to which a patented invention is technically disclosed through public code poses a non-trivial challenge: patent descriptions are written in technical prose, while code contributions are executable instructions. Comparing them requires methods that can bridge these differences. We address this challenge using sentence embedding models, neural networks that transform text into dense numerical vectors where semantically related content occupies nearby points in embedding space (Vaswani et al., 2017; Devlin et al., 2019; Reimers and Gurevych, 2019). These methods extend the tradition of textual analysis (Gentzkow et al., 2019) by capturing conceptual relationships even when documents share no common words. We use these measures to construct a digital disclosure index that reflects the semantic proximity between a patent's description and the public code contributions made by its assignee on the project(s) mentioned in the patent document.

We find that patents accompanied by high levels of digital disclosure, defined as those whose descriptions exhibit above-median semantic similarity to publicly released code, attract citations from inventors located significantly farther away than matched controls, with an estimated increase of approximately 17%, or around 770 kilometers. Our findings are robust to alternative distance measures, similarity thresholds, and sentence embedding models, as well as to controls for patent quality, novelty, and economic value.

An alternative interpretation of our results is that broader spatial diffusion of citations reflects 'permissionless innovation' rather than better technical understanding Thierer (2016); Chesbrough

[1] These organizations represent the 26.6% of all patents granted and the 48% of all public code contributions to these projects over the period. See Petralia (2024).

and Van Alstyne (2015). Patents referencing public code repositories often rely on permissively licensed code, granting users explicit rights to use and redistribute the technology without negotiation, which may reduce legal frictions for distant adopters. We address this concern by showing that citation distances are similar regardless of license type, and that within-patent variation in the timing of code contributions predicts citation distance even when license terms remain unchanged.

Our study contributes to several related strands of literature. First, we contribute to the large body of research studying the geography of knowledge spillovers, which has documented that knowledge flows decline sharply with geographic distance (Audretsch and Feldman, 1996; Jaffe et al., 1993; Varga, 2000; Paci and Usai, 2000; Autant-Bernard, 2001; Breschi and Lissoni, 2001; Thompson and Fox-Kean, 2005; Autant-Bernard et al., 2007; Feldman and Kogler, 2010; Murata et al., 2014; Ganguli et al., 2020; Boschma, 2005). This literature identifies social networks, professional relationships, labor mobility, and the nature of knowledge itself as important drivers of localization (Agrawal et al., 2008; Sorenson et al., 2006; Diemer and Regan, 2022; Cotterlaz and Guillouzouic, 2025). We complement this literature by identifying a mechanism through which the spatial frictions constraining knowledge diffusion can be reduced. Our findings show that the distance over which knowledge diffuses depends partly on the form in which it is disclosed.

This mechanism is particularly relevant for understanding the persistence of geographic concentration in the digital era. Although digital technologies substantially reduce communication and transmission costs, their effects on distant knowledge flows remain conditional on shared expertise and absorptive capacity (Forman and van Zeebroeck, 2019). The effect of distance on patent-citation flows has remained remarkably stable despite the rise of the internet (Cotterlaz and Guillouzouic, 2025), while remote teams remain less effective at combining tacit knowledge to produce breakthrough ideas (Lin et al., 2023). Consistent with these findings, innovative activity remains concentrated in a relatively small number of regions (Feldman and Kogler, 2010; Balland et al., 2020). This concentration is especially strong for complex activities and software (Forman and Goldfarb, 2020; Wachs et al., 2022; Chattergoon and Kerr, 2022). Our findings help bridge this apparent tension by showing that digital availability alone may be insufficient to reduce spatial frictions. Digital technologies reduce the cost of transmitting knowledge, but the ability of knowledge to diffuse across space also depends on whether it is disclosed in a form that distant inventors can understand, test, and apply.

More broadly, our results have implications for the spatial distribution of innovative opportunities. Because localized knowledge spillovers disproportionately benefit regions that already possess substantial innovative capabilities, they can reinforce cumulative advantages and contribute to persistent regional inequalities (Balland et al., 2020; Moretti, 2021; Kemeny et al., 2022; Petralia et al., 2023; Pinheiro et al., 2025).

We also add to the growing body of research on disclosure and innovation (Furman et al., 2021; Hegde and Luo, 2018; Hegde et al., 2023; Baruffaldi and Simeth, 2020; Lück et al., 2020; Chondrakis et al., 2021; Berkes and Nencka, 2024; Gross, 2023; Saidi and Žaldokas, 2021; Beyhaghi et al., 2023; Baruffaldi et al., 2024). Prior work shows that earlier disclosure, machine translation, and intermediate reusable disclosures can increase access to and reuse of technical knowledge (Baruffaldi and Simeth, 2020; Büttner et al., 2022; Boudreau and Lakhani, 2015). We complement this literature by studying how the form of disclosure shapes its ability to transfer knowledge over space. Our findings highlight the limitations of relying solely on legal texts to diffuse knowledge (Seymore, 2009; Fromer, 2008; Rai, 1999).

Lastly, our results also speak to the classic distinction between tacit and codified knowledge (Polanyi, 1966; Cowan et al., 2000; Gertler, 2003; Howells, 1996; Balconi, 2002), since executable code represents a particularly actionable form of knowledge codification: it forces precision, reduces ambiguity, and allows experimentation. By showing that code-backed disclosure diffuses over greater distances, we provide evidence that actionable codification can reduce, without eliminating, inventors' dependence on geographically localized tacit knowledge.

The remainder of the article proceeds as follows. Section 2 describes the patent and public-code data, the construction of the matched sample, and our measure of actionable digital disclosure. Section 3 presents the main estimates and robustness analyses, while Section 4 discusses the implications of our findings.

2 Data Sources

This study draws data from two primary sources. Information about organizations' contributions to OSS projects comes from Petralia (2024)[2], while details about their patenting activity in the United States (US) originate from the United States Patent and Trademark Office (USPTO).

Petralia (2024) links two decades of contributions (2005-2024) to 98 of the most popular OSS projects in GitHub, the most popular code-hosting service for OSS development[3], with patent

[2]Available at: https://doi.org/10.7910/DVN/UQNVHF

[3]GitHub has reported a user base of over 100 million developers as of 2023: https://github.blog/2023-01-25-100-million-developers-and-counting/.

filings in the United States for 1556 organizations. These organizations represent the 26.6% of all patents granted and the 48% of all OSS contributions to these projects over the period.

While GitHub hosts the most comprehensive collection of code repositories, spanning diverse programming languages, frameworks, and libraries across domains such as web development, machine learning, data analysis, and mobile applications, not all repositories are equally valuable. The OSS projects in Petralia (2024) were selected based on three quality metrics: forks, watchers, and commits. Forks indicate how many users have copied a repository to modify or extend it; watchers reflect the number of users actively following a repository for updates; and commits measure the volume of contributions a project has received. Together, these metrics capture project popularity, reuse potential, community engagement, perceived value, and ongoing maintenance activity. The dataset includes all projects that ranked among the top 100 in at least one of these metrics.

This database provides detailed information about each contribution, including the author's name and email, submission timestamp, and metrics such as the number of files changed and lines added or deleted. Contributors are attributed to organizations using the web domain of their email addresses, with organizational information retrieved from Compustat[4] and the Whois database, supplemented by manual matching where automated lookups were insufficient. Tables A.5, A.6, and A.7 summarize the types of projects included, the most actively contributed repositories, and the organizations making those contributions. For a detailed description of the database construction and the methodology linking organizations' patenting and GitHub activities, see Appendix A of Petralia (2025).

From our second data source, the USPTO, we extract patent metadata for the patenting activity of these organizations, including filing and grant dates, citations, inventor and assignee locations, assigned technological classifications, and organizational types, and the full text of the patent document. We include only citation records of utility patents filed between January 1, 2005, and December 31, 2024. We also restrict to citations categorized as 'cited by applicant' and made by assignees classified as corporations, whether domestic or foreign. Self-citations, where the citing and cited patents share the same assignee, and citations in which the citing patent was filed before the cited patent are excluded. Finally, we retain only those records for which both the citing and cited patents can be geolocated. These restrictions allow us to focus on formal knowledge flows between external inventors and to measure the distance over which those flows occur.

[4] See: https://wrds-www.wharton.upenn.edu/pages/get-data/compustat-capital-iq-standard-poors/compustat/

We exclude self-citations to focus on knowledge flows to external inventors, which are more directly affected by disclosure quality and accessibility. Including self-citations would limit our ability to measure diffusion to outside inventors, as these often reflect strategic patenting or internal knowledge management (Kuhn et al., 2020). Similarly, we discard citations in which the citing patent was filed before the cited patent, as these cannot represent forward-looking knowledge flows (Arora et al., 2018). This is because applicants should have had access to the cited knowledge at the time of invention, and any chronological inconsistency would undermine the interpretation of citations as evidence of knowledge transmission. Finally, as Kuhn et al. (2020) argue, applicant-added citations provide a more reliable indication of knowledge transmission than examiner-added citations because they reflect prior knowledge identified by the inventors themselves.

We extend Petralia (2025) by extracting the actual source code from each contribution and linking it to USPTO patents held by the same firm. We extract all code additions, the lines of code newly introduced in each contribution, by cloning each repository and parsing its complete commit history. This allows us to capture the actual source code that developers contributed, not just summary statistics. We then measure the semantic similarity between patents that mention these project(s) and the code additions made by the same firms, thus measuring the degree to which a firm's contributions relate to the technical content of its patents.

To measure similarity between a firm's patented inventions and its public code contributions, we use a two-step approach. First, we use neural language models to transform text into dense numerical vectors (embeddings) that encode semantic meaning (Vaswani et al., 2017; Devlin et al., 2019). This methodology is increasingly adopted in economics and finance research (Gentzkow et al., 2019) because it moves beyond keyword-based matching to capture nuanced semantic relationships. These models are trained on vast corpora of text and code to learn representations where semantically similar content occupies nearby points in vector space. Importantly, embeddings capture conceptual similarities even when documents share no words, making them well-suited for comparing patent descriptions with source code.

For our baseline analysis, we use DeepSeek-Coder-6.7B (Guo et al., 2024), an open-source language model trained on 2 trillion tokens (87% source code across 87 programming languages and 13% natural language). This training mix allows the model to generate embeddings that meaningfully represent both the prose of patent descriptions and the syntax of programming languages. We extract embeddings from the model's final hidden layer to obtain a 4,096-dimensional vector for each patent description and code contribution.

Given two embedding vectors **p** (patent) and **c** (code), we compute cosine similarity:

$$\mathrm{sim}(\mathbf{p}, \mathbf{c}) = \frac{\mathbf{p} \cdot \mathbf{c}}{|\mathbf{p}||\mathbf{c}|}$$

This metric measures the cosine of the angle between two vectors, yielding values between -1 and 1, with higher values indicating greater semantic similarity. Cosine similarity is standard in NLP because it is invariant to vector magnitude, focusing on direction rather than length (Reimers and Gurevych, 2019). See Appendix A.3 for more details.

2.1 Sample Construction: Patents Referencing Public Code and Matched Control Group

To construct a sample of patents that are more likely to have disclosed part of their technical content online, along with a comparable control group, we begin by identifying all patents that explicitly mention any of the OSS projects included in this study. To do so we rely on the keyword-based identification strategy proposed in Petralia (2025), detailed in Tables A.8 and A.9. Table A.10 lists the ten OSS projects most frequently mentioned in patents between 2005 and 2024.

We focus on US-based patents, defined as those assigned to US corporations with at least one US-based inventor. Each focal patent is paired (without replacement) with a control patent that does not reference any of the OSS projects used in this study, shares at least one Cooperative Patent Classification (CPC) group, and matches the focal patent in the number of unique inventors and assignees. Control patents are selected from the pool of US patents filed within six months of the focal patent, with priority given to minimizing differences in filing date, followed by grant date. This approach closely follows the matching strategies employed by Jaffe et al. (1993); Thompson and Fox-Kean (2005); Murata et al. (2014); Ganguli et al. (2020), but differs in that those studies typically analyze citations to focal patents relative to a randomly selected comparison group.

Table 1 summarizes the matched sample of focal and control patents. The final number of matched patents is 33,789 out of 35,839 cases (94.3%). The first panel reports key outcome variables. On average, focal patents received slightly more citations (18.7) than control patents (17), with both distributions being highly skewed. The descriptive spatial patterns also differ between focal and control patents. While the distance between citing and cited inventor(s) is relatively high in both groups, focal patents tend to be cited from farther away despite receiving a similar overall number of citations. This initial pattern motivates our analysis of whether the availability of related public code accounts for differences in the spatial distribution of citations.

Table 1: Descriptive Statistics

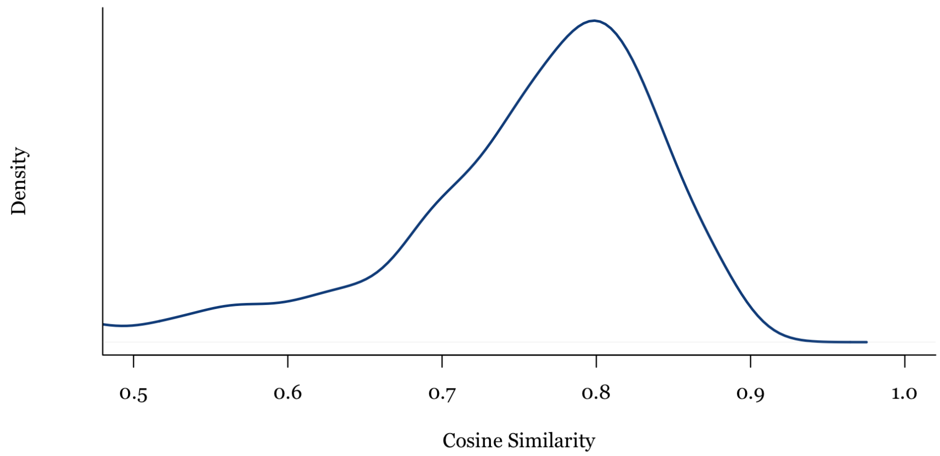
	Focal Patents			Control Patents		
Variables	Mean	Median	SD	Mean	Median	SD
Citations	18.7	5	46.3	17	4	44.4
Patent-level Mean Citation Distance (km)	4454	3778.7	3031.6	4265	3523.3	3044.5
Patent-level Median Citation Distance (km)	4132.2	3319.5	3653.4	3987.3	2932.9	3617.1
Matching Level Stats:	Mean	Median	SD			
Filing Date Difference (days)	3.1	0	13.2			
Granting Date Difference (Days)	354.1	210	407.5			
Number of Matched Patents: 33789 of 35839						

Figure 1 shows the distribution of semantic similarity scores between patent descriptions and code contributions. Using the DeepSeek-Coder-6.7B model (Guo et al., 2024), we compute the cosine similarity between each patent description and all code contributions made by the assignee firm in the repository mentioned in the patent, from six months before filing through five years after publication (approximately 142 million pairwise comparisons).

We use this distribution to classify patents by their level of 'digital disclosure' at the time of publication. Specifically, we restrict our analysis to code contributions made between six months prior to the patent's filing date and its publication date. For each patent, we identify the maximum similarity score observed within this period. We then divide the distribution of these maximum scores into quartiles, classifying patents in the bottom two quartiles as having moderate digital disclosure and those in the top two quartiles (above the median) as having high digital disclosure. Patents meeting this threshold indicate cases where inventors have access to a higher level of technical disclosure through the firm's public contribution activity.

Of the 33,789 focal patents in our sample, 22,840 (67.6%) have at least one code contribution by the assignee in the mentioned project during the relevant period. Figure 1 displays the distribution of the maximum similarity scores for these patents.

Figure 1: Distribution of Pre-Publication Maximum Similarity Scores



Notes: Kernel density estimate of similarity scores between patent descriptions and their most similar code contribution by the assignee firm in the mentioned repository, from six months before patent filing to the publication date.

Further details on the data and methodology are provided in Appendix A.2 and A.3, while the data and replication code are publicly available and described in Appendix A.1.[5]

[5] See: https://zenodo.org/records/18790146?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjdjZGM4NjM2LTBkZDktNGM5YS04ODZjLTU2YWE5ZmMwMjNmOSIsImRhdGEiOnt9LCJyYW5kb20iOiIwOTAzNjA2OTkzOGRlYzkyODEzNmU
8wDZFe9-3B3PDYfKzrcGUaaaItKE1J_dLe8C4pHLWfiHBc3HT7hvlRamWC2B178-8uuIV9v4VqqgVr5w5JN8Ag

3 Results

The primary goal of this section is to evaluate how the availability of detailed technical disclosure, in the form of publicly available digital code, affects the spatial diffusion of knowledge. To do so we compare patents that rely on public digital code infrastructure to closely matched counterfactual patents that do not. We consider this approach effective because technologies built on this infrastructure often require contributors to engage directly with the codebase (Petralia, 2025; Henkel, 2004, 2006), leading to a wealth of technical disclosure that is rarely available elsewhere. As firms modify and extend this infrastructure, they create publicly accessible and highly codified technical knowledge that offers a clear and unambiguous representation of some of the invention's technical features, in an environment that facilitates experimentation, prototyping, and testing. In addition, from the perspective of follow-on innovators, variation in the level of public digital disclosure across pre-existing technologies is largely outside the control of individual firms, as they typically have limited influence over competitors' disclosure decisions.

It is unclear in principle how the availability of detailed technical disclosure may affect the geography of follow-on innovation. On the one hand, by facilitating learning and experimentation, it could enable broader participation and ultimately more geographically dispersed innovation (Wachs et al., 2022; Meszaros and Wachs, 2024). On the other hand, its impact may be limited: disclosure could constrain experimentation and narrow technological search (Boudreau and Lakhani, 2015), benefits may concentrate among firms with prior capabilities or close connections to the focal innovator (Wright et al., 2023; Fang et al., 2024), and complex licensing or infringement concerns may deter new entrants, reducing incentives for innovation.

The main specification, shown in Equation 1, evaluates whether there are systematic differences in the spatial distribution of citations between focal patents and their matched counterparts. Specifically, D_i measures the patent-level mean distance: we first calculate the average distance between citing and cited inventors separately for each cited patent, so that every patent receives equal weight in the analysis. The key explanatory variable, Focal_i, is a binary indicator that equals one if patent i mentions any of the projects described in Section A.2, and zero otherwise. The coefficient of interest, β_1, thus captures whether citations to focal patents originate on average from greater distances. We control for the overall citation intensity of each patent (Citations_i), the number of claims (Claims_i), and include matched set fixed effects, $\delta_{g(i)}$, where $g(i)$ indexes the matched group to which patent i belongs. The error term ϵ_i captures idiosyncratic variation, and standard errors are clustered at the matched set level, as in Cameron et al. (2012).

$$D_i = \beta_1 \cdot \text{Focal}_i + \beta_2 \cdot \log(\text{Citations})_i + \beta_3 \cdot \log(\text{Claims})_i + \delta_{g(i)} + \epsilon_i \tag{1}$$

Column (1) of Table 2 reports regression estimates of how spatial diffusion patterns differ between focal and control patents in terms of average geographic proximity between citing and cited inventors. Focal patents receive citations that systematically originate from locations that are farther away. Results in Column (1) naively assume that patents referencing these projects are associated with higher levels of technical disclosure. Although firms leveraging publicly available digital infrastructure often generate accessible codified information, not all firms actively contribute to this process. Many rely on existing components without contributing back, which can limit the extent of technical disclosure captured in the patents. Moreover, mentions of these projects in patents could reflect minimal or incidental use, legal requirements, or strategic signaling, rather than active contribution. To address these concerns, columns (2) through (4) assess whether differences in spatial diffusion are concentrated among patents accompanied by related public code.

To evaluate the extent to which a patented invention is disclosed in publicly available code repositories, we measure the semantic proximity between the patent description and code contributions made by the patent assignee to the project(s) referenced in the patent. We include contributions from six months before patent filing through to the publication date.

For each patent description, we compute its cosine similarity with all available code contributions during this time window and identify the maximum similarity score, which we use as our patent-level measure of digital disclosure. We then categorize focal patents into three groups: *no disclosure* for patents for which we found no code contributions; *moderate disclosure* for patents with maximum similarity below the median; and *high disclosure* for patents with maximum similarity above the median (see Appendix A.3 for a detailed description of the procedure, including examples).

Table 2 reports the results of this disaggregated analysis. A particularly interesting case is Column (2), which can be seen as a placebo test: it focuses on patents with no code contributions. For this group, focal and control patents exhibit no statistically significant difference in citation distance. By contrast, a clear spatial difference appears in Columns (3) and (4). Focal patents are, on average, cited by inventors located significantly farther away than those of the control group, and this difference is more evident in the subsample of patents with high digital disclosure (Column

4), which shows a distance premium of approximately 17%. Given that the average inventor-to-inventor distance in this sample is 4,492 km, this implies an increase of approximately 770 km. This pattern is consistent with actionable technical disclosure reducing distant inventors' reliance on geographically localized knowledge.

Table 2: Difference in the Spatial Footprint of Citations to Focal and Control Patents for Different Levels of Digital Disclosure

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.064***	0.007	0.037**	0.159***
	(0.009)	(0.015)	(0.016)	(0.014)
Citations (in logs)	−0.0002	0.012	−0.011	0.001
	(0.005)	(0.008)	(0.009)	(0.008)
Claims (in logs)	−0.036***	−0.052***	−0.032**	−0.011
	(0.008)	(0.014)	(0.015)	(0.013)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	67,578	24,572	21,730	21,276
Adjusted R^2	0.031	0.027	0.038	0.019

*p<0.1; **p<0.05; ***p<0.01

Notes: Errors are clustered at the level of the matching pair.

We perform a series of robustness checks to ensure that our results are not driven by the choice of distance measure (Tables A.12 to A.13) or the threshold used to identify patents with high digital disclosure (Table A.14). As Kuhn et al. (2020) note, the choice between using filing dates or granting dates in citation analyses can influence measures of citation timing, technological similarity, and spillovers. Since our sample construction relies on filing dates, it could inherit biases driven by these date-based dynamics. To address this concern, Table A.16 restricts the analysis to matched patent pairs with granting dates within 180 days of each other, showing results remain consistent.

Another potential concern is that our matching procedure does not account for the geographic location of inventors, which could introduce bias if focal and control patents differ systematically in their spatial composition. For example, observed differences in diffusion patterns might simply reflect pre-existing geographic factors rather than digital disclosure. To address this, we perform a robustness check that replicates Table 2 with an additional matching constraint: we require that at least one inventor share the same city of residence between each focal and control patent pair. As shown in Table A.20, the results remain consistent.

A related concern is that the observed results might be driven primarily by large technology firms, which dominate both open-source contributions and patent filings in our sample. We address this issue by replicating our main analysis restricting the sample to citations originating from small and medium-sized firms, defined as those with fewer than 500 patents over the entire study period (Table A.21). This definition also includes patents without an assignee, which typically represent individual inventors or small entities. The results suggest that actionable digital disclosure may be particularly valuable for smaller organizations, for which informal networks and geographic proximity to knowledge sources might otherwise be more important. This finding is in line with Gross (2023), who shows that the costs of restricted disclosure fall disproportionately on smaller and more technologically focused firms, which have fewer existing alternatives and less flexibility to redirect their R&D efforts.

The robustness of the results extends to including additional patent-level controls such as measures of breakthroughness (Kelly et al., 2021), novelty (Arts et al., 2021), and economic value (Kogan et al., 2017) (Tables A.17 to A.18). These controls address concerns that our findings might be driven by heterogeneity in patent quality or importance rather than digital disclosure. Breakthrough patents may show wider spatial diffusion because they represent fundamental technological shifts that appeal to diverse geographic areas, while highly novel patents might show more concentrated spatial patterns due to greater barriers to adoption. Similarly, economically valuable patents may attract attention from a broader set of firms and locations, accelerating their spatial diffusion independent of code availability. Crucially, if high digital disclosure correlates with these patent characteristics, our baseline results could reflect differences in patent quality rather than the association with related public code.

We also assess whether our results could be driven by routine or incidental code contributions unrelated to the firm's innovative activities. Table A.22 replicates our analysis using only commits authored by patent inventors, identified by matching contributor names to inventors listed on patents filed by the same organization. These contributions are more likely to reflect technically substantive, innovation-oriented work rather than maintenance tasks or administrative updates.

Finally, we address the concern that our results might be driven by the geographic sorting of open source developers rather than by the disclosure of technical information. We replicate our baseline analysis excluding all citations from firms that never contributed to any of the open source projects in our sample. By focusing on citations from non-contributing firms, we limit the extent to which the estimates reflect direct community membership or participation in the same public-code networks. As shown in Table A.23, our results remain robust, suggesting that the observed patterns are driven by the availability of actionable technical disclosure rather than by shared participation in open source communities.

3.1 Permissionless Innovation as an Alternative Channel

The robustness checks discussed before address concerns related to measurement, matching, and patent heterogeneity. However, one alternative explanation survives these tests and requires separate treatment: the possibility that higher geographic diffusion reflects reduced legal frictions.

Patents with high digital disclosure often incorporate public code distributed under permissive licenses such as Apache or MIT. These licenses function as standardized global contracts, granting downstream users explicit rights to use, modify, and redistribute the underlying code without negotiation. This creates a 'permissionless innovation' channel, through which distant inventors cite the patent not because the disclosed code clarifies technical content, but because it lowers the transaction costs of adoption.

Consider an inventor evaluating whether to build on a US patent. Absent clear licensing terms, geographical distance compounds uncertainty since distant actors face higher costs in assessing litigation risk, establishing licensing agreements, or leveraging informal networks. When the patent relies heavily on permissively licensed code, however, a substantial portion of the technology becomes available under terms that are both transparent and globally enforceable. The license may reduce the distance penalty associated with legal negotiation. If this mechanism explains our results, the observed citation patterns would reflect lower contractual barriers rather than more actionable technical disclosure.

To distinguish actionable disclosure from the permissionless innovation channel, we follow two complementary strategies. The first leverages heterogeneity in license type among the project(s) referenced by focal patents.

Not all open-source licenses are equal with respect to patent rights. The Apache 2.0 license is particularly relevant because it explicitly addresses patent rights: users receive automatic, royalty-free rights to any contributor patents necessary to use the code, along with provisions that terminate those rights for users who initiate patent litigation. These provisions make Apache 2.0 the clearest case of permissionless innovation as it allows unrestricted use and redistribution while explicitly resolving the patent uncertainty.

Table 3 reports results from estimating our baseline specification including matched sets for which the focal patent has high digital disclosure (Column4 of Table 2), augmented with an interaction between the focal patent indicator and a dummy for whether the referenced project(s) is/are distributed under a permissive license (Apache 2.0). Of the 10,638 focal patents in this subsample, 838 (7.9%) reference permissively licensed projects. The results show that the increase in citation distance is similar for patents linked to permissive licenses and those that are not, which is inconsistent with the idea that permissionless innovation is driving our baseline results.

Table 3: Baseline Results with License Type Interaction (All Patents with Disclosure)

	Average Distance (in logs)
	High Disclosure
Focal	0.163***
	(0.015)
Citations (in logs)	0.002
	(0.008)
Claims (in logs)	−0.011
	(0.013)
Focal * Permissive	−0.046
	(0.053)
Period	2005-2024
Fixed Effects	Matched Set
Observations	21,276
Adjusted R^2	0.019

*p<0.1; **p<0.05; ***p<0.01

Notes: Errors are clustered at the level of the matching pair.

Our second strategy takes a different approach by holding license type constant and relying on variation in the timing of digital disclosure. We focus on the subsample of patents for which we do not observe any digital disclosure at the time of publication (Column 2 of Table 2). For these patents, the licensing terms associated with the project(s) they mention remained unchanged throughout our observation period and only the availability of detailed technical disclosure varies across citations, depending on when they take place. Because license type is fixed within each patent, changes in citation distance associated with the timing of code contributions are difficult to attribute to changes in legal frictions.

The shift in focus from patents to citation requires a different matching strategy. Rather than matching focal and control patents, we now match individual citations. For each citation received by a focal patent, we identify a control patent closely following the criteria in Jaffe et al. (1993) and Thompson and Fox-Kean (2005). Specifically, each focal citation is matched to a control patent that shares at least one CPC group, has the same number of inventors, and the same number of assignees. These control patents are randomly drawn (with replacement) from the pool of US patents filed within 30 days of the focal citing patent.[6]

[6] We restrict our sample to citations filed within 10 years of the focal patent's filing date. Since we track code contributions for up to 5 years after filing, this window allows sufficient time for knowledge diffusion while ensuring the link between disclosure and citation remains relevant.

Table 4: Difference in the Spatial Footprint of Focal and Control Citations

	Average Distance (in logs)			
	All	High Disclosure	No Disclosure	No Disclosure
	(1)	(2)	(3)	(4)
Focal Citation	-0.529^{***}	-0.386^{***}	-0.624^{***}	-0.624^{***}
	(0.003)	(0.005)	(0.005)	(0.005)
Focal Citation * Exposure (Similarity)			0.136^{***}	
			(0.013)	
Focal Citation * Exposure (Dummy)				0.110^{***}
				(0.011)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
	Filing Lag	Filing Lag	Filing Lag	Filing Lag
Observations	1,136,908	263,794	463,646	463,646
Adjusted R^2	0.133	0.134	0.124	0.124

*p<0.1; **p<0.05; ***p<0.01

Notes: All specifications include matched-set fixed effects (at the level of the cited-citing pair and its matched control) and filing lag fixed effects (in years). The latter account for the mechanical relationship between citation timing and exposure: citations filed later are more likely to have been exposed to higher levels of digital disclosure and to originate from more distant locations as technologies diffuse. Errors are clustered at the matched-set level.

Table 4 presents estimates from the citation-level matching design. Column (1) pools all citations and serves as a baseline check: citations made to focal patents originate from inventors located significantly closer to the cited inventors. This pattern aligns with the literature on localized knowledge spillovers (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014; Ganguli et al., 2020) and with recent evidence on the geographic concentration of software patenting (Chattergoon and Kerr, 2022; Forman and Goldfarb, 2020).

Column (2) replicates this analysis for citations to patents with high digital disclosure at publication. The localization effect, while still negative and significant, is notably smaller in magnitude, which is consistent with our main results: patents accompanied by detailed code disclosure attract citations from more distant inventors, reducing the typical home bias observed in patent citations.

Columns (3) and (4) provide the core test for the permissionless innovation channel. We focus on patents that had no meaningful digital disclosure at publication and interact the focal citation indicator with two different measures of subsequent exposure to digital disclosure: a continuous measure of the maximum semantic similarity between the patent description and the code available at the time of citation (Column 3), and a binary indicator for whether any relevant code contribution existed at that time (Column 4). In both cases, we find that citations exposed to higher levels of digital disclosure are significantly less spatially concentrated. This pattern is difficult to reconcile with the permissionless innovation hypothesis. License terms did not change over time, only the availability of detailed technical disclosure. If legal frictions drove our baseline results, we would not expect citation distance to respond to the timing and intensity of code disclosure. These results are robust to restricting the sample to citations filed within five years of the focal patent, alternative clustering strategies, and cited-citing year fixed effects (Tables A.24-A.26).

4 Concluding Remarks

Knowledge flows remain strongly localized even as technical information becomes increasingly available online. This paper identifies actionable disclosure as one mechanism capable of reducing the spatial frictions that constrain knowledge diffusion. In the context of the patent system, publicly available executable code allows distant inventors to inspect, test, and build upon technical knowledge that may remain difficult to use when disclosed only through patent documents.

Using a novel dataset linking two decades of patent filings with public code contributions to major open-source software repositories, we find that patents accompanied by high levels of digital disclosure attract citations from inventors located approximately 17 percent (770 km) farther away than matched control patents. This difference is concentrated among patents whose assignees contributed code that is semantically similar to the patented invention, a pattern consistent with improved technical understanding. Our findings are robust to alternative distance measures, matching strategies that account for geographic sorting, and controls for patent quality, novelty, and economic value. We also provide evidence against an alternative explanation based on permissionless innovation: the difference in citation distance is similar across license types, and within-patent variation in the timing of code contributions predicts citation distance even when licensing terms remain unchanged.

These findings suggest that the challenge of effective disclosure is not solely about timing or availability, but about form. Patent documents are usually written in legal prose and technical jargon, and while this satisfies disclosure requirements, it introduces frictions that make the underlying knowledge more difficult to transfer across geographic boundaries. Executable code, by contrast, forces precision, eliminates ambiguity, and enables experimentation, qualities that reduce these frictions and facilitate knowledge transfer across geographic boundaries.

The results also speak to the persistence of geographic concentration in the digital era. Lower transmission costs and global access to information do not necessarily eliminate the advantages of proximity when distant inventors still face substantial costs of interpretation and application. By reducing these costs, actionable disclosure may weaken one mechanism through which innovative opportunities accumulate in established technological centers.

We focus on software-related technologies, where executable code provides a natural complement to patent disclosure. Whether similar patterns can be found in other technological domains, such as biotechnology or materials, remains an open question. Additionally, while our identification strategy addresses concerns about selection and confounding, we cannot entirely rule out unobserved heterogeneity between patents with high and low digital disclosure. Finally, our measure of digital disclosure only captures semantic similarity between patents and code contributions, which may not fully reflect the practical utility of the disclosed information for inventors.

Future work could explore whether these patterns extend to other innovation outcomes, such as licensing, startup formation, or technology adoption speed. A particularly interesting direction would be to study whether diffusion to more distant inventors translates into new innovative capabilities outside established technological centers, and how actionable disclosure shapes not just the location but also the direction of subsequent innovation. More broadly, investigating how the form of disclosure interacts with other channels through which firms communicate their innovative capabilities, such as scientific publications or open data, could shed light on how knowledge transfer can be more effectively designed across different industries and institutional contexts. Our findings may also be informative beyond the software domain. The mechanism we document, that knowledge diffuses more broadly when it is actionable, rests on the ability of inventors to prototype, test, and experiment with disclosed inventions. Artificial intelligence is rapidly making knowledge more actionable by extending these capabilities to other fields, lowering the cost of experimentation and enabling the recombination of existing knowledge at scale. If so, the patterns we observe for open code may offer preliminary insight into the potential effects of artificial intelligence on knowledge diffusion.

Data and Code Availability

The data and code used in this study are publicly available at https://zenodo.org/records/18790146?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjdjZGM4NjM2LTBkZDktNGM5YS04ODZjLTU2YWE5
8wDZFe9-3B3PDYfKzrcGUaaaItKE1J_dLe8C4pHLWfiHBc3HT7hvlRamWC2B178-8uuIV9v4VqqgVr5w5JN8Ag.

References

Agrawal, Ajay, Devesh Kapur, and John McHale, "How do spatial and social proximity influence knowledge flows? Evidence from patent data," *Journal of Urban Economics*, 2008, *64* (2), 258–269.

Almeida, Paul and Bruce Kogut, "Localization of knowledge and the mobility of engineers in regional networks," *Management science*, 1999, *45* (7), 905–917.

Andrews, Michael, "Bar Talk: Informal Social Networks, Alcohol Prohibition, and Invention," *Alcohol Prohibition, and Invention (June 25, 2023)*, 2023.

Arora, Ashish, Sharon Belenzon, and Honggi Lee, "Reversed citations and the localization of knowledge spillovers," *Journal of Economic Geography*, 2018, *18* (3), 495–521.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research policy*, 2021, *50* (2), 104144.

Audretsch, David B and Maryann P Feldman, "R&D spillovers and the geography of innovation and production," *The American economic review*, 1996, *86* (3), 630–640.

Autant-Bernard, Corinne, "The geography of knowledge spillovers and technological proximity," *Economics of Innovation and New Technology*, 2001, *10* (4), 237–254.

_ , Jacques Mairesse, and Nadine Massard, "Spatial knowledge diffusion through collaborative networks," *Papers in Regional Science*, 2007, *86* (3), 341–350.

Balconi, Margherita, "Tacitness, codification of technological knowledge and the organisation of industry," *Research Policy*, 2002, *31* (3), 357–379.

Balland, Pierre-Alexandre, Cristian Jara-Figueroa, Sergio G Petralia, Mathieu PA Steijn, David L Rigby, and César A Hidalgo, "Complex economic activities concentrate in large cities," *Nature human behaviour*, 2020, *4* (3), 248–254.

Baruffaldi, Stefano H and Markus Simeth, "Patents and knowledge diffusion: The effect of early disclosure," *Research Policy*, 2020, *49* (4), 103927.

Baruffaldi, Stefano, Markus Simeth, and David Wehrheim, "Asymmetric information and R&D disclosure: Evidence from scientific publications," *Management Science*, 2024, *70* (2), 1052–1069.

Berkes, Enrico and Peter Nencka, "Knowledge access: The effects of carnegie libraries on innovation," *Review of Economics and Statistics*, 2024, pp. 1–45.

Bessen, James, "Patents and the diffusion of technical information," *Economics letters*, 2005, *86* (1), 121–128.

Beyhaghi, Mehdi, Pooyan Khashabi, and Ali Mohammadi, "Pre-grant patent disclosure and analyst forecast accuracy," *Management Science*, 2023, *69* (5), 3140–3155.

Boschma, Ron, "Proximity and innovation: a critical assessment," *Regional studies*, 2005, *39* (1), 61–74.

Boudreau, Kevin J. and Karim R. Lakhani, "'Open' Disclosure of Innovations, Incentives and Follow-on Reuse: Theory on Processes of Cumulative Innovation and a Field Experiment in Computational Biology," *Research Policy*, 2015, *44*, 4–19.

Breschi, Stefano and Francesco Lissoni, "Knowledge spillovers and local innovation systems: a critical survey," *Industrial and Corporate Change*, 2001, *10* (4), 975–1005.

_ **and** _ , "Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows," *Journal of economic geography*, 2009, *9* (4), 439–468.

Büttner, Benjamin, Murat Firat, and Emilio Raiteri, "Patents and knowledge diffusion: The impact of machine translation," *Research Policy*, 2022, *51* (10), 104584.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 2012.

Chattergoon, Brad and William R Kerr, "Winner takes all? Tech clusters, population centers, and the spatial transformation of US invention," *Research Policy*, 2022, *51* (2), 104418.

Chesbrough, Henry and Marshall Van Alstyne, "Permissionless innovation," *Communications of the ACM*, 2015, *58* (8), 24–26.

Chondrakis, George, Carlos J Serrano, and Rosemarie H Ziedonis, "Information disclosure and the market for acquiring technology companies," *Strategic Management Journal*, 2021, *42* (5), 1024–1053.

Cotterlaz, Pierre and Arthur Guillouzouic, "The percolation of knowledge across space," *Journal of International Economics*, 2025, *153*, 104026.

Cowan, Robin, Paul A David, and Dominique Foray, "The explicit economics of knowledge codification and tacitness," *Industrial and corporate change*, 2000, *9* (2), 211–253.

de Rassenfosse, Gaétan, "Patent rights and cumulative innovation," *The Patentist Living Literature Review*, 2025. Available at SSRN 5983416.

_ , Gabriele Pellegrino, and Emilio Raiteri, "Do patents enable disclosure? Evidence from the invention secrecy act," *International Journal of Industrial Organization*, 2024, *92*, 103044.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in "Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)" 2019, pp. 4171–4186.

Diemer, Andreas and Tanner Regan, "No inventor is an island: Social connectedness and the geography of knowledge flows in the US," *Research Policy*, 2022, *51* (2), 104416.

Fang, Hongbo, Patrick Park, James Evans, James Herbsleb, and Bogdan Vasilescu, "Weak Ties Explain Open Source Innovation," 2024.

Feldman, Maryann P and Dieter F Kogler, "Stylized facts in the geography of innovation," *Handbook of the Economics of Innovation*, 2010, *1*, 381–410.

Forman, Chris and Avi Goldfarb, "Concentration and Agglomeration of IT Innovation and Entrepreneurship: Evidence from Patenting," Technical Report w27338, National Bureau of Economic Research 2020.

_ and Nicolas van Zeebroeck, "Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?," *Research Policy*, 2019, *48* (8), 103697.

Fromer, Jeanne C, "Patent disclosure," *Iowa L. Rev.*, 2008, *94*, 539.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger, "Disclosure and subsequent innovation: Evidence from the patent depository library program," *American Economic Journal: Economic Policy*, 2021, *13* (4), 239–270.

Ganguli, Ina, Jeffrey Lin, and Nicholas Reynolds, "The paper trail of knowledge spillovers:

evidence from patent interferences," *American Economic Journal: Applied Economics*, 2020, *12* (2), 278–302.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gertler, Meric S, "Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there)," *Journal of economic geography*, 2003, *3* (1), 75–99.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, Karin Hoisl, Christian Le Bas, Alessandra Luzzi, Laura Magazzini, Lionel Nesta, Önder Nomaler, Neus Palomeras, Pari Patel, Marzia Romanelli, and Bart Verspagen, "Inventors and invention processes in Europe: Results from the PatVal-EU survey," *Research Policy*, 2007, *36* (8), 1107–1127.

Gross, Daniel P, *The consequences of invention secrecy: Evidence from the USPTO patent secrecy program in World War II*, National Bureau of Economic Research Cambridge, Massachusetts, USA, 2019.

_ , "The hidden costs of securing innovation: The manifold impacts of compulsory invention secrecy," *Management Science*, 2023, *69* (4), 2318–2338.

Guo, Daya, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li et al., "DeepSeek-Coder: When the Large Language Model Meets Programming–The Rise of Code Intelligence," *arXiv preprint arXiv:2401.14196*, 2024.

Hegde, Deepak and Hong Luo, "Patent publication and the market for ideas," *Management Science*, 2018, *64* (2), 652–672.

_ **, Kyle Herkenhoff, and Chenqi Zhu**, "Patent publication and innovation," *Journal of Political Economy*, 2023, *131* (7), 1845–1903.

Henkel, Joachim, "The jukebox mode of innovation-a model of commercial open source development," *Available at SSRN 578142*, 2004.

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

Howells, Jeremy, "Tacit knowledge," *Technology analysis & strategic management*, 1996, *8* (2), 91–106.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson, "Geographic localization of knowledge spillovers as evidenced by patent citations," *the Quarterly journal of Economics*, 1993, *108* (3), 577–598.

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

Kemeny, Tom, Sergio Petralia, and Michael Storper, "Disruptive innovation and spatial inequality," *Regional Studies*, 2022, pp. 1–18.

Klevorick, Alvin K, Richard C Levin, Richard R Nelson, and Sidney G Winter, "On the sources and significance of interindustry differences in technological opportunities," *Research policy*, 1995, *24* (2), 185–205.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological innovation, resource allocation, and growth," *The quarterly journal of economics*, 2017, *132* (2), 665–712.

Kuhn, Jeffrey, Kenneth Younge, and Alan Marco, "Patent citations reexamined," *The RAND Journal of Economics*, 2020, *51* (1), 109–132.

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang, "Towards General Text Embeddings with Multi-stage Contrastive Learning," *arXiv preprint arXiv:2308.03281*, 2023.

Lin, Yiling, Carl Benedikt Frey, and Lingfei Wu, "Remote collaboration fuses fewer breakthrough ideas," *Nature*, 2023, *623*, 987–991.

Lück, Sonja, Benjamin Balsmeier, Florian Seliger, and Lee Fleming, "Early disclosure of invention and reduced duplication: An empirical test," *Management Science*, 2020, *66* (6), 2677–2685.

Machlup, Fritz and Edith Penrose, "The patent controversy in the nineteenth century," *The Journal of Economic History*, 1950, *10* (1), 1–29.

Meszaros, Gabor and Johannes Wachs, "The Dynamics of Innovation in Open Source Software

Ecosystems," *arXiv preprint arXiv:2411.14894*, 2024.

Moretti, Enrico, "The effect of high-tech clusters on the productivity of top inventors," *American Economic Review*, 2021, *111* (10), 3328–3375.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura, "Localized knowledge spillovers and patent citations: A distance-based approach," *Review of Economics and Statistics*, 2014, *96* (5), 967–985.

Murray, Fiona and Siobhán O'Mahony, "Exploring the foundations of cumulative innovation: Implications for organization science," *Organization science*, 2007, *18* (6), 1006–1021.

OECD, *Patents, Innovation and Economic Performance: OECD Conference Proceedings*, Paris, France: OECD Publishing, 2004.

Ouellette, Lisa Larrimore, "Do patents disclose useful information," *Harv. JL & Tech.*, 2011, *25*, 545.

Paci, Raffaele and Stefano Usai, "Externalities, knowledge spillovers and the spatial distribution of innovation," *GeoJournal*, 2000, *49*, 381–390.

Petralia, Sergio, "Data Repository for "Open Source Software as Digital Platforms to Innovate"," 2024.

_ , "Open source software as digital platforms to innovate," *Journal of Economic Behavior & Organization*, 2025, *237*, 107109.

_ , Thomas Kemeny, and Michael Storper, "The transformative effects of tacit technological knowledge," III Working Paper 103, International Inequalities Institute, London School of Economics and Political Science 2023.

Pinheiro, Flavio L., Pierre-Alexandre Balland, Ron Boschma, and Dominik Hartmann, "The dark side of the geography of innovation: relatedness, complexity and regional inequality in Europe," *Regional Studies*, 2025, *59* (1).

Polanyi, Michael, *The tacit dimension*, University of Chicago press, 1966.

Rai, Arti Kaur, "Regulating scientific research: Intellectual property rights and the norms of science," *Nw. UL Rev.*, 1999, *94*, 77.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese

BERT-Networks," in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics 11 2019.

Saidi, Farzad and Alminas Žaldokas, "How does firms' innovation disclosure affect their banking relationships?," *Management Science*, 2021, *67* (2), 742–768.

Sampat, Bhaven N, "A survey of empirical evidence on patents and innovation," Technical Report, National Bureau of Economic Research 2018.

Seymore, Sean B, "The teaching function of patents," *Notre Dame L. Rev.*, 2009, *85*, 621.

Singh, Jasjit, "Collaborative networks as determinants of knowledge diffusion patterns," *Management science*, 2005, *51* (5), 756–770.

Sorenson, Olav, Jan W. Rivkin, and Lee Fleming, "Complexity, networks and knowledge flow," *Research Policy*, 2006, *35* (7), 994–1017.

Thierer, Adam, *Permissionless innovation: The continuing case for comprehensive technological freedom*, Mercatus Center at George Mason University, 2016.

Thompson, Peter and Melanie Fox-Kean, "Patent citations and the geography of knowledge spillovers: A reassessment," *American Economic Review*, 2005, *95* (1), 450–460.

Varga, Attila, "Local academic knowledge spillovers and the concentration of economic activity," *Journal of Regional Science*, 2000, *40* (2), 289–309.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017, *30.*

Wachs, Johannes, Mariusz Nitecki, William Schueller, and Axel Polleres, "The Geography of Open Source Software: Evidence from GitHub," *Technological Forecasting and Social Change*, 2022, *176*, 121478.

Wright, Nataliya Langburd, Frank Nagle, and Shane Greenstein, "Open source software and global entrepreneurship," *Research Policy*, 2023, *52* (9), 104846.

Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang et al., "Qwen2 Technical Report," *arXiv preprint arXiv:2407.10671*, 2024.

A Appendix

A.1 Data Repository Descriptor

This section describes the datasets available in the data repository, which contains four files to replicate the main results presented in Section 3. The repository also includes R and python scripts to reproduce most tables and show how to embed patent text and code.

A.1.1 Patent-to-Code Similarity Data (Pats2Code.parquet)

This file contains pairwise similarity scores between patent descriptions and code contributions, computed using the DeepSeek-Coder-6.7B neural language model. Each row represents a patent-commit pair with its corresponding cosine similarity score. The file is stored in Apache Parquet format for efficient storage and querying.

Table A.1: Variable Definitions: Pats2Code.parquet

Variable	**Description**
patent_id	Unique USPTO patent identifier
commit	Commit hash identifying the code contribution as it appears in Petralia (2024)
similarity	Cosine similarity between patent description embedding and code contribution embedding (range: -1 to 1)

A.1.2 Patent-Level Regression Data (MainS3.csv)

This file contains patent-level data for the matched sample analysis presented in Table 2. Each row represents a patent (either focal or control) with its citation patterns and digital disclosure measures.

Table A.2: Variable Definitions: MainS3.csv

Variable	Description
MatchID	Unique identifier for the focal-control patent pair
patent_id	Unique USPTO patent identifier
Citations	Total number of forward citations received by the patent
MeanDistance	Mean geographic distance (km) between citing and cited inventors
MedianDistance	Median geographic distance (km) between citing and cited inventors
Focal	Binary indicator: 1 if patent references an OSS project, 0 if control
IsPermissive	Binary indicator: TRUE if the referenced project(s) use a permissive license, FALSE otherwise
Repos	Repository name(s) referenced by the patent
baseline_max_sim	Maximum similarity score between patent description and code contributions (DeepSeek-6.7B)
baseline_commit	GitHub commit hash corresponding to the maximum similarity score
baseline_quartile	Quartile of the maximum similarity score distribution
NClaims	Number of claims in the patent

A.1.3 Citation-Level Regression Data (MainS3.1.csv)

This file contains citation-level data for the analysis presented in Table 4. Each row represents a unique citation event with information about geographic distance and exposure to digital disclosure.

Table A.3: Variable Definitions: MainS3.1.csv

Variable	Description
UniqueCitation	Unique identifier for the citation event (matched-set)
patent_id	USPTO patent identifier of the cited patent
Focal	Binary indicator: 1 if citation is to a focal patent, 0 if to a control
Exposure	Maximum similarity score between patent description and code contributions at the time of citation
MeanDistance	Mean geographic distance (km) between citing and cited inventors
MedianDistance	Median geographic distance (km) between citing and cited inventors
NoDisclosure	Binary indicator: 1 if the cited patent had no meaningful digital disclosure at publication
HighDisclosure	Binary indicator: 1 if the cited patent had high digital disclosure at publication
YearCited	Filing year of the cited patent
YearCiting	Filing year of the citing patent
FLag	Filing lag in years between citing and cited patents

A.1.4 License Data (LicenseData.csv)

This file contains license information for each repository included in the study. Each row represents a repository with its corresponding license identifier and type classification.

Table A.4: Variable Definitions: LicenseData.csv

Variable	Description
Repo	Repository name in the format owner/repository
LicenseID	Standard license identifier (e.g., Apache-2.0, GPL-3.0, MIT)
LicenseType	License classification: Permissive or Restrictive

A.2 Supporting Material for Section 2

Table A.5: Summary of Repository Types

Functionality	Projects
Communication	8
Compute Resources	33
Data Management and Storage	19
Integration	22
Monitoring and Observability	8
Orchestration	8

Notes: This table outlines the types of OSS projects included in the dataset, categorized by their core functionality. **Compute resources** provide the environments and tools needed to run applications, including operating systems, virtualization, and cloud services. **Data Management and Storage** involve organizing, protecting, and preserving data. **Communication** infrastructures enable information exchange between components via protocols and networking layers. **Orchestration** automates the deployment, scaling, and updating of software. **Integration** tools connect disparate systems through APIs and adapters. Finally, **Monitoring and Observability** systems track performance and reliability by analyzing telemetry data such as logs, metrics, and traces.

Table A.6: Top Repositories by Number of Contributions (Commits)

	OSSProject	Description	FirstYear	Commits
1	Linux	a family of kernels powering a wide range of devices from servers to phones	2005	$1,329,904$
2	Android OS	a software platform that runs smartphones and tablets	2008	$1,051,600$
3	Liferay	a portal platform for developing corporate intranets and digital experiences	2006	$848,005$
4	Mozilla	a community behind web browsers and open web standards	2005	$808,603$
5	Llvm	a collection of tools for compiling and optimizing code in many languages	2005	$505,935$
6	Webkit engine	a browser engine powering many mobile and desktop web browsers	2005	$281,077$
7	Percona cluster	a solution to run MySQL databases across many servers	2008	$238,261$
8	Percona server	an enhanced MySQL-compatible database with extra reliability features	2008	$226,644$
9	Percona xtrabackup	a tool for creating hot backups of MySQL databases	2008	$215,040$
10	FreeBSD	a UNIX-like operating system known for stability and advanced networking	2005	$185,945$

Notes: This table, taken from Petralia (2025), presents the top 10 projects, along with a concise description of them generated using GPT-4. The column 'First Commit' indicates the year of the first commit in the sample, while the column 'Commits' provides information on the number of commits for each project until 2024.

Table A.7: Top Domains by Number of Commits

	Web Domain	OSS Project	Commits	File Changes
1	google.com	Android OS	946,803	2,266,292
2	liferay.com	Liferay	799,646	4,326,602
3	mozilla.com	Mozilla	390,577	6,765,051
4	oracle.com	Percona xtrabackup	174,434	1,615,859
5	oracle.com	Percona cluster	174,219	1,527,424
6	oracle.com	Percona server	170,935	1,508,064
7	apple.com	Webkit engine	143,714	2,008,083
8	intel.com	Linux	105,043	248,318
9	oracle.com	MySQL	92,244	2,547,732
10	google.com	Tensorflow	91,358	897,648
11	redhat.com	Ceph	90,537	366,540
12	apple.com	Llvm	78,252	315,974
13	oracle.com	Oracle graal	76,377	979,509
14	google.com	Llvm	66,072	325,840
15	redhat.com	Linux	62,632	160,511

Notes: This table, taken from Petralia (2025), shows the top 15 web domains (ranked by number of commits) in terms of contributions per project. The columns 'Commits' and 'File Changes' provide information on the number of commits and file changes each domain has submitted for each project until 2024.

Table A.8: List of Keywords to Search (1/2)

Repository	Keyword
ansible/ansible	ansible
aosp-mirror/platform_frameworks_base	android os
apache/airflow	apache airflow
apache/camel	apache camel
apache/dubbo	dubbo
apache/flink	apache flink
apache/kafka	apache kafka
apache/nuttx	nuttx
apache/rocketmq	rocketmq
apache/spark	apache spark
apache/superset	apache superset
ArduPilot/ardupilot	ardupilot
bitcoin/bitcoin	bitcoin
BVLC/caffe	caffe framework
caddyserver/caddy	caddyserver
CartoDB/cartodb	cartodb
ceph/ceph	ceph
chartjs/Chart.js	chart js
cockroachdb/cockroach	cockroachdb
duckdb/duckdb	duckdb
elastic/elasticsearch	elasticsearch
esp8266/Arduino	esp8266
ethereum/go-ethereum	ethereum
facebook/hhvm	hhvm
facebook/react-native	react native
fastapi/fastapi	fastapi
freebsd/freebsd-src	freebsd
godotengine/godot	godot engine
google/skia	skia
gradle/gradle	gradle
grafana/grafana	grafana
grpc/grpc	grpc
hashicorp/terraform-provider-aws	terraform
helm/charts	helm charts
jhipster/generator-jhipster	jhipster
keras-team/keras	keras
Kitware/CMake	cmake
Kitware/ParaView	paraview
kubernetes/kubernetes	kubernetes
laravel/framework	laravel
libretro/RetroArch	retroarch
liferay/liferay-portal	liferay
llvm/llvm-project	llvm
lobehub/lobe-chat	lobe chat
lodash/lodash	lodash
louislam/uptime-kuma	uptime kuma
MariaDB/server	mariadb
mermaid-js/mermaid	mermaid js
microsoft/PowerToys	powertoys

Table A.9: List of Keywords to Search (2/2)

Repository	Keyword
mongodb/mongo	mongodb
mozilla/gecko-dev	mozilla
mrdoob/three.js	mrdoob
mybatis/mybatis-3	mybatis
mysql/mysql-server	mysql
neo4j/neo4j	neo4j
nestjs/nest	nest js
netdata/netdata	netdata
nodejs/node	node js
nomic-ai/gpt4all	gpt4all
ollama/ollama	ollama
open-webui/open-webui	open webui
openai/whisper	openai whisper
opencv/opencv	opencv
openjdk/jdk	openjdk
opensearch-project/OpenSearch	opensearch
opf/openproject	openproject
oracle/graal	oracle graal
OSGeo/gdal	osgeo gdal
percona/percona-server	percona server
percona/percona-xtrabackup	percona xtrabackup
percona/percona-xtradb-cluster	percona cluster
postgres/postgres	postgresql
protocolbuffers/protobuf	protobuf
pytorch/pytorch	pytorch
qemu/qemu	qemu
rabbitmq/rabbitmq-server	rabbitmq
ravendb/ravendb	ravendb
reactos/reactos	reactos
reduxjs/redux	redux js
SchedMD/slurm	slurm
shadcn-ui/ui	shadcn ui
shadowsocks/shadowsocks	shadowsocks
Significant-Gravitas/AutoGPT	autogpt
supabase/supabase	supabase
sveltejs/svelte	svelte js
syncthing/syncthing	syncthing
tailwindlabs/tailwindcss	tailwindcss
taosdata/TDengine	tdengine
tensorflow/tensorflow	tensorflow
torvalds/linux	linux
trilinos/Trilinos	trilinos
typicode/json-server	typicode
vitejs/vite	vite js
vuejs/vue	vue js
WebKit/WebKit	webkit engine
wireshark/wireshark	wireshark
zephyrproject-rtos/zephyr	zephyr os
zulip/zulip	zullip

Table A.10: Top Projects in Terms of Mentions In Patent Documents

Repository	Mentions
linux	69,537
mozilla	7,903
mysql	6,847
kubernetes	3,385
freebsd	3,156
mongodb	2,979
bitcoin	2,850
tensorflow	2,821
opencv	2,590
postgresql	1,837

Notes: This table shows the top 10 projects in terms of the number of USPTO patents between 2005 and 2024 that mentions them.

A.3 Patent to Code Similarity Measures

We use a two-step approach to measure the semantic similarity between a firm's patented inventions and its public code contributions. This methodology leverages neural language models to move beyond traditional keyword-based matching, allowing us to capture nuanced conceptual relationships even when patent descriptions and source code share no common words.

The first step transforms textual content, both patent descriptions and code contributions, into dense numerical vectors known as embeddings. These representations encode semantic meaning in a high-dimensional space where conceptually similar content occupies nearby points (Vaswani et al., 2017; Devlin et al., 2019). This embedding-based approach has gained widespread adoption in economics and finance (Gentzkow et al., 2019) because it captures semantic relationships that surface-level text matching cannot detect. For instance, a patent describing a 'distributed consensus algorithm' and code implementing a 'blockchain validation function' may share deep conceptual similarity despite differing vocabulary.

We compute embeddings using three different neural language models that span a range of complexity and architectural design choices. This multi-model approach allows us to assess the robustness of our findings across different model capacities, context lengths, and architectural designs.

For our baseline analysis, we use DeepSeek-Coder-6.7B (Guo et al., 2024), an open-source decoder-only Transformer model specifically designed for code intelligence tasks. We select this model as our baseline because it offers the best balance between semantic precision and computational feasibility for our corpus of over 142 million patent-code comparisons. The model was trained from scratch on 2 trillion tokens consisting of 87% source code (spanning 87 programming languages) and 13% natural language (including English code-related text from GitHub Markdown and StackExchange). This training composition enables the model to generate embeddings that meaningfully represent both the prose of patent descriptions and the syntax of programming code.

DeepSeek-Coder-6.7B contains 6.7 billion parameters organized in a 32-layer decoder-only Transformer architecture. The model supports a context window of 16,384 tokens using Rotary Position Embeddings (RoPE), allowing it to process lengthy patent descriptions and code without truncation for virtually all code and patents. We extract the final hidden layer representations and apply mean pooling across non-padding tokens to obtain a single embedding vector and obtain a 4,096-dimensional representation for each document. The model is publicly available at `https://huggingface.co/deepseek-ai/deepseek-coder-6.7b-base`.

As a first robustness check, we evaluate DeepSeek-Coder-1.3B, a smaller variant from the same model family (Guo et al., 2024). This model shares the same training data composition and architectural design principles as the 6.7B variant but with a reduced parameter count of 1.3 billion parameters organized in a 24-layer architecture. It produces 2,048-dimensional embeddings compared to the baseline's 4,096 dimensions. This comparison allows us to assess whether the higher precision and richer representations of the larger model are necessary for detecting meaningful patent-to-code relationships, or whether a more compact model capturing broader semantic patterns is sufficient (available at `https://huggingface.co/deepseek-ai/deepseek-coder-1.3b-base`).

As a second robustness check, we employ GTE-Qwen2-7B-instruct (Li et al., 2023), an embedding model built on the Qwen2 architecture (Yang et al., 2024). This model differs from our baseline in three important respects. First, it supports a substantially larger context window of 32,768 tokens, double that of DeepSeek-Coder-6.7B, enabling the processing of longer patent descriptions and code contributions without truncation. Second, while DeepSeek models are decoder-only architectures trained primarily for code generation through next-token prediction, GTE-Qwen2 is explicitly optimized for producing high-quality text embeddings through contrastive learning objectives. Third, unlike the DeepSeek models which were trained predominantly on source code (87%), GTE-Qwen2 is a general-purpose embedding model trained on broader corpora that includes code but is not specialized for it. While this means GTE-Qwen2 may lack some of the fine-grained code syntax understanding that domain-specific models have, it provides a valuable test of whether our findings depend on specialized code intelligence or whether they hold with more general semantic representations. The model contains approximately 7 billion parameters and produces 3,584-dimensional embeddings. Since this is an instruction-tuned embedding model, we prepend a task-specific instruction to patent descriptions: "Given a patent description, retrieve related code implementations." The model is publicly available at https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct.

Table A.11 summarizes the key differences across these three embedding models. The practical trade-offs between these models extend beyond computational considerations such as memory requirements and processing time. Larger models with more parameters can, in principle, learn finer-grained distinctions in meaning. For example, the 6.7B parameter baseline may better differentiate between code that implements a 'breadth-first search' versus a 'depth-first search', recognizing that while both are graph algorithms, they have distinct computational properties and use cases. A smaller model like DeepSeek-1.3B might represent both as simply 'graph search algorithms', collapsing important technical distinctions.

However, this additional nuances in interpretation are not unambiguously beneficial for the aim of our study. When measuring whether public contributions relate to patented inventions, we are often interested in broad conceptual similarity, does this code implement ideas from the same technological domain as the patent? rather than fine implementation details. A smaller model that captures higher-level patterns (e.g., 'machine learning classifier' or 'network protocol handler') may be equally or more informative for detecting knowledge spillovers than a larger model that distinguishes between specific algorithmic variants. In other words, the 'right' level of abstraction depends on whether patent-to-code spillovers operate at the level of general concepts or specific implementations.

Context length represents another important distinction. While the DeepSeek models can process up to 16,384 tokens, GTE-Qwen2-7B extends this to 32,768 tokens. For patents with extensive descriptions or code contributions spanning hundreds of lines, longer context windows may capture relevant content that would otherwise be truncated. At the same time, longer contexts introduce additional noise, and it remains an empirical question whether the extended window improves similarity detection for our application.

As a second step, and given two embedding vectors $\mathbf{p}$ (patent) and $\mathbf{c}$ (code), we compute the

Table A.11: Comparison of Embedding Models

Characteristic	**DeepSeek-1.3B**	**DeepSeek-6.7B**	**GTE-Qwen2-7B**
Parameters	1.3B	6.7B (Baseline)	7B
Architecture	Decoder-only	Decoder-only	Decoder-only (Qwen2)
Training Objective	Next-token + FIM	Next-token + FIM	Contrastive learning
Code in Training	87% (87 languages)	87% (87 languages)	General + code
Context Window	16,384 tokens	16,384 tokens	32,768 tokens
Embedding Dimension	2,048	4,096	3,584
Robustness Check	Lower precision	—	Larger context

cosine similarity between them:

$$\mathrm{sim}(\mathbf{p}, \mathbf{c}) = \frac{\mathbf{p} \cdot \mathbf{c}}{|\mathbf{p}||\mathbf{c}|}$$

This metric measures the cosine of the angle between two vectors (ranging from -1 to 1), with higher values indicating greater semantic similarity. Cosine similarity is standard in NLP applications because it is invariant to vector magnitude, focusing on directional alignment rather than absolute length (Reimers and Gurevych, 2019). This property is particularly important when comparing documents of varying lengths.

As an example, consider the following excerpt from patent 10,789,544, filed by Google on April 5, 2016[7], which describes a method for batching inputs to a machine learning model, alongside a small extract of the most semantically similar code contributed by Google shortly before the patent filing to the TensorFlow repository (explicitly mentioned in the patent)[8]. The patent extract (the abstract for simplicity) describes the high-level method: receiving a stream of input tensors, queuing them, and generating batched tensors either when enough inputs accumulate or when latency constraints are met. The code excerpt implements this logic at a functional level by creating a queue (*FIFOQueue* in line 14) to hold the input tensors for batch processing. For example, the patent's description of 'adding the respective input tensor from each request to a queue' corresponds directly to the creation of a queue (q in line 14) and the invocation of *enqueue_many* (line 18) to add input tensors into it. In this example, the sentence embedding model captures the underlying meaning of both the patent description and the code snippet by converting them into numerical vectors that reflect their semantic content, despite differences in format and language, by focusing on shared concepts such as input batching, queuing, and processing.

[7] See https://patents.google.com/patent/US10789544

[8] See https://github.com/tensorflow/tensorflow/commit/aba52f4b4b405d15a37b7c56b026a514b9022b45

We interpret this similarity score as a measure of the degree of digital technical disclosure of the patent: a higher similarity implies that the substance of the patent is more likely to have been made available through public code.

Figure A.1: Extract from patent 10,789,544 describing a method to 'batch inputs to a machine learning model' filed by Google on April 5,2016

Abstract:

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for batching inputs to machine learning models. One of the methods includes receiving a stream of requests, each request identifying a respective input for processing by a first machine learning model; adding the respective input from each request to a first queue of inputs for processing by the first machine learning model; determining, at a first time, that a count of inputs in the first queue as of the first time equals or exceeds a maximum batch size and, in response: generating a first batched input from the inputs in the queue as of the first time so that a count of inputs in the first batched input equals the maximum batch size, and providing the first batched input for processing by the first machine learning model.

Figure A.2: Extract From Commit **aba52f4b4b4** submitted by Google on March 18, 2016

```python
@@input_producer
def input_producer(input_tensor, element_shape=None, num_epochs=None,
                   shuffle=True, seed=None, capacity=32, shared_name=None,
                   summary_name=None, name=None):
  with ops.op_scope([input_tensor], name, "input_producer"):
    input_tensor = ops.convert_to_tensor(input_tensor, name="input_tensor")
    element_shape = input_tensor.get_shape()[1:].merge_with(element_shape)
    if not element_shape.is_fully_defined():
      raise ValueError("Either `input_tensor` must have a fully defined shape "
                       "or `element_shape` must be specified")
    if shuffle:
      input_tensor = random_ops.random_shuffle(input_tensor, seed=seed)
    input_tensor = limit_epochs(input_tensor, num_epochs)
    q = data_flow_ops.FIFOQueue(capacity=capacity,
                                dtypes=[input_tensor.dtype.base_dtype],
                                shapes=[element_shape],
                                shared_name=shared_name, name=name)
    enq = q.enqueue_many([input_tensor])
    queue_runner.add_queue_runner(queue_runner.QueueRunner(q, [enq]))
    if summary_name is not None:
      logging_ops.scalar_summary("queue/%s/%s" % (q.name, summary_name),
                                 math_ops.cast(q.size(), dtypes.float32) *
                                 (1. / capacity))
    return q
```

A.4 Supporting Tables & Robustness Checks

Table A.12: Replication of Table 2 Using Median Distance

	Median Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.048^{***}	-0.047^{**}	-0.037	0.257^{***}
	(0.013)	(0.022)	(0.025)	(0.021)
Citations (in logs)	-0.131^{***}	-0.112^{***}	-0.167^{***}	-0.104^{***}
	(0.007)	(0.012)	(0.014)	(0.012)
Claims (in logs)	-0.034^{***}	-0.069^{***}	-0.018	0.018
	(0.013)	(0.022)	(0.024)	(0.021)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	67,578	24,572	21,730	21,276
Adjusted R^2	0.032	0.015	0.053	0.026

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2 using the median geographic distance between citing and cited inventors. Errors are clustered at the level of the matching pair.

Table A.13: Replication of Table 2 Using Differences in City of Residence of Inventors

	Different City (Share)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.002^{***}	-0.001	-0.001	0.010^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Citations (in logs)	-0.002^{***}	-0.001^{**}	-0.004^{***}	0.00002
	(0.0004)	(0.001)	(0.001)	(0.001)
Claims (in logs)	-0.003^{***}	-0.004^{***}	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	67,578	24,572	21,730	21,276
Adjusted R^2	0.053	0.019	0.088	0.023

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2, but instead of measuring geographic distance between citing and cited inventors, it uses the share of citing-cited inventor pairs that are located in different cities, where cities are considered as the smallest location identifier provided by the USPTO. Errors are clustered at the level of the matching pair.

Table A.14: Replication of Table 2 By Similarity Quartile

	Mean Distance (in logs)				
	No Disclosure	Q1	Q2	Q3	Q4
	(1)	(2)	(3)	(4)	(5)
Focal	0.007	0.039*	0.038	0.145***	0.174***
	(0.015)	(0.022)	(0.024)	(0.020)	(0.020)
Citations (in logs)	0.012	0.039***	−0.058***	−0.002	0.005
	(0.008)	(0.013)	(0.013)	(0.011)	(0.011)
Claims (in logs)	−0.052***	−0.060***	−0.006	0.012	−0.035*
	(0.014)	(0.023)	(0.021)	(0.019)	(0.019)
Period	2005-2024	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set	Matched Set
Observations	24,572	10,854	10,876	10,818	10,458
Adjusted R^2	0.027	0.037	0.044	0.021	0.018

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2 by similarity quartile. Errors are clustered at the level of the matching pair.

Table A.15: Replication of Table 2 Using Different Sentence Embedding Models

	Moderate Disclosure			High Disclosure		
	(1)	(2)	(3)	(4)	(5)	(6)
	Robustness	Baseline	Robustness	Robustness	Baseline	Robustness
Focal	0.068***	0.037**	0.067***	0.127***	0.159***	0.123***
Citations (in logs)	-0.012	-0.011	0.009	0.001	0.001	-0.022**
Claims (in logs)	-0.026*	-0.032**	-0.023	-0.019	-0.011	-0.028*
Model	DeepSeek 1.3B	DeepSeek 6.7B	GTE-Qwen2-7B	DeepSeek 1.3B	DeepSeek 6.7B	GTE-Qwen2-7B
Role	Lower precision	Baseline	Larger context	Lower precision	Baseline	Larger context
Parameters	1.3B	6.7B	7B	1.3B	6.7B	7B
Context Length	16K tokens	16K tokens	32K tokens	16K tokens	16K tokens	32K tokens
Embedding Dimension	2048	4096	3584	2048	4096	3584
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set	Matched Set	Matched Set

This table replicates the analysis in Table 2 using three sentence embedding models. DeepSeek-Coder-6.7B is our baseline model, chosen for its balance of semantic precision and computational feasibility. DeepSeek-Coder-1.3B provides a robustness check with reduced parameter count and lower-dimensional embeddings (2,048 vs. 4,096), testing whether smaller models capturing broader patterns yield similar results. GTE-Qwen2-7B-instruct provides a robustness check with a different architecture (Qwen2-based) and an extended context window (32,768 tokens vs. 16,384), testing sensitivity to model design and context length. Columns (1)–(3) show results for moderate disclosure and columns (4)–(6) for high disclosure. All specifications include matched-pair fixed effects. Standard errors are clustered at the matched-pair level.

Table A.16: Replication of Table 2 Keeping Only Matches Where Differences in Granting Dates Are Less Than a Year

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.057^{***}	0.007	0.040^{**}	0.140^{***}
	(0.011)	(0.019)	(0.020)	(0.018)
Citations (in logs)	0.002	0.013	−0.010	0.006
	(0.006)	(0.010)	(0.011)	(0.010)
Claims (in logs)	-0.025^{**}	-0.047^{***}	−0.013	−0.001
	(0.010)	(0.017)	(0.019)	(0.017)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	45,030	16,508	14,800	13,722
Adjusted R^2	0.030	0.023	0.043	0.015

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Notes: This table replicates the analysis from Table 2 keeping only matches where differences in granting dates between the focal and control patent is less than 365 days. Errors are clustered at the level of the matching pair.

Table A.17: Replication of Table 2 Including a Patent Measure of Breakthrough Innovation From Kelly et al. (2021)

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.052^{***}	-0.025	0.037	0.125^{***}
	(0.012)	(0.022)	(0.023)	(0.017)
Citations (in logs)	0.008	0.003	0.022^{*}	0.009
	(0.006)	(0.011)	(0.013)	(0.010)
Claims (in logs)	-0.053^{***}	-0.072^{***}	-0.040^{**}	-0.038^{**}
	(0.010)	(0.019)	(0.020)	(0.016)
Breakthrough Indicator	-0.069^{***}	-0.050	-0.058	-0.074^{***}
	(0.019)	(0.035)	(0.037)	(0.028)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	29,294	9,338	8,008	11,948
Adjusted R^2	0.028	0.027	0.046	0.011

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Notes: This specification incorporates the breakthrough indicator developed by Kelly et al. (2021) as a control variable to account for the potential influence of groundbreaking innovations on the geographic diffusion of patented knowledge. This index identifies patents that are both novel, textually dissimilar from prior patents, and impactful, meaning they are closely aligned with subsequent innovations, using forward and backward similarity ratios. Breakthrough patents are defined as those in the upper tail of this importance distribution (top 10%). The sample size differs from the baseline because matched pairs are excluded if either the focal or control patent lacks the breakthrough indicator. Errors are clustered at the matching-pair level.

Table A.18: Replication of Table 2 Including a Patent Measure of Economic Value From Kogan et al. (2017)

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.031***	−0.011	−0.026	0.113***
	(0.012)	(0.023)	(0.021)	(0.017)
Citations (in logs)	−0.014**	−0.002	−0.032**	−0.003
	(0.007)	(0.013)	(0.014)	(0.011)
Claims (in logs)	−0.028**	−0.037*	−0.005	−0.030*
	(0.011)	(0.022)	(0.021)	(0.017)
Economic Value	−0.001***	−0.001**	−0.001	−0.001**
	(0.0003)	(0.001)	(0.0005)	(0.001)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	36,818	10,298	12,436	14,084
Adjusted R^2	0.027	0.018	0.038	0.018

*p<0.1; **p<0.05; ***p<0.01

Notes: This specification includes the economic value measure developed by Kogan et al. (2017) as a control variable to account for the potential influence of a patent's economic significance on the geographic diffusion of ideas. This measure estimates the private economic value of a patent by analyzing stock market reactions to patent grants, providing a market-based assessment of the patent's importance. The number of patents in this specification differs from the baseline because matched pairs are dropped if either the focal or control patent lacks economic value information. Errors are clustered at the matching pair level.

Table A.19: Replication of Table 2 Including a Patent Measure of Novelty From Arts et al. (2021)

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.062^{***}	-0.009	0.056^{***}	0.133^{***}
	(0.011)	(0.019)	(0.020)	(0.016)
Citations (in logs)	0.001	0.001	0.014	-0.002
	(0.006)	(0.011)	(0.012)	(0.010)
Claims (in logs)	-0.051^{***}	-0.070^{***}	-0.044^{**}	-0.032^{**}
	(0.010)	(0.017)	(0.018)	(0.015)
Cosine Similarity	-0.467^{***}	-0.535^{*}	-0.604^{**}	-0.333
	(0.149)	(0.277)	(0.297)	(0.214)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	37,352	12,412	10,700	14,240
Adjusted R^2	0.026	0.020	0.045	0.011

*p<0.1; **p<0.05; ***p<0.01

Notes: This specification includes the cosine similarity measure developed by Arts et al. (2021) to control for the effect that patent novelty may have on the geographic diffusion of ideas. This metric quantifies the semantic similarity between a patent and a corpus of prior patents, where lower similarity scores indicate greater novelty. It captures the extent of innovation introduced by the patent relative to existing technologies. The sample size differs from the baseline because matched pairs are excluded if either the focal or control patent lacks cosine similarity data. Errors are clustered at the matching pair level.

Table A.20: Replication of Table 2 with City-Level Matching

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.058***	0.006	0.075***	0.095***
	(0.011)	(0.019)	(0.021)	(0.017)
Citations (in logs)	0.0001	0.035***	−0.027**	−0.013
	(0.006)	(0.010)	(0.011)	(0.009)
Claims (in logs)	0.021*	0.026	0.026	0.014
	(0.011)	(0.019)	(0.020)	(0.018)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	43,284	15,758	13,764	13,762
Adjusted R^2	0.038	0.034	0.033	0.037

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2 with an additional matching constraint: we require that at least one inventor share the same city of residence between each focal and control patent pair. Cities are the smallest location identifier provided by the USPTO. Errors are clustered at the level of the matching pair.

Table A.21: Replication of Table 2 Including Only Citations From Small Firms

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.094***	0.024	0.071**	0.230***
	(0.016)	(0.025)	(0.032)	(0.029)
Citations (in logs)	−0.013	0.014	−0.045***	−0.011
	(0.009)	(0.014)	(0.016)	(0.015)
Claims (in logs)	−0.052***	−0.064***	−0.060**	−0.006
	(0.015)	(0.024)	(0.028)	(0.027)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	25,236	10,688	7,510	7,038
Adjusted R^2	0.045	0.018	0.093	0.019

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2 but restricts the sample to citations originating from small and medium-sized firms, defined as those with fewer than 500 patents over the entire study period (2005–2024). This threshold captures a broad range of firms outside the largest patent holders, including patents without an assignee, which typically represent individual inventors or small entities. Errors are clustered at the level of the matching pair.

Table A.22: Replication of Table 2 Using Only Public Contributions by Inventors

	Average Distance (in logs)	
	Moderate Disclosure	High Disclosure
	(1)	(2)
Focal	0.075^{***}	0.141^{***}
	(0.017)	(0.015)
Citations (in logs)	-0.024^{**}	-0.002
	(0.010)	(0.008)
Claims (in logs)	-0.019	-0.019
	(0.017)	(0.014)
Period	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set
Observations	18,962	18,544
Adjusted R^2	0.035	0.018

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the analysis from Table 2 but computes digital disclosure using only code contributions authored by patent inventors (they represent approximately 28% of all patent to commit comparison instances). Contributors are matched to inventors by first and last name when using the same organizational email domain. This restriction ensures that the measured disclosure reflects technically substantive contributions closely tied to the firm's innovative activities rather than routine maintenance or incidental commits. Errors are clustered at the level of the matching pair.

Table A.23: Replication of Table 2 Considering Only Citations from Non Contributors

	Average Distance (in logs)			
	All	No Disclosure	Moderate Disclosure	High Disclosure
	(1)	(2)	(3)	(4)
Focal	0.075^{***}	0.028^{*}	0.049^{***}	0.161^{***}
	(0.009)	(0.015)	(0.017)	(0.015)
Citations (in logs)	0.008^{*}	0.018^{**}	-0.001	0.012
	(0.005)	(0.008)	(0.010)	(0.009)
Claims (in logs)	-0.034^{***}	-0.050^{***}	-0.027^{*}	-0.010
	(0.009)	(0.014)	(0.016)	(0.014)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
Observations	62,974	23,024	20,322	19,628
Adjusted R^2	0.029	0.024	0.034	0.023

*p<0.1; **p<0.05; ***p<0.01

Notes: This table replicates the baseline analysis from Table 2 but restricts the sample to citations originating from assignees that have never contributed to any of the open source projects in our dataset. This robustness check addresses the concern that our results might be driven by the specific geographic distribution or clustering of active open source participants. By focusing on non-contributing firms, we limit the extent to which the estimates reflect direct community membership or participation in the same public-code networks. Matched sets are excluded if either the focal or control patent received no citations from non-contributing assignees. Errors are clustered at the level of the matching pair.

Table A.24: Replication of Table 4 Using a 5 Year Window

	Average Distance (in logs)			
	All	High Disclosure	No Disclosure	No Disclosure
	(1)	(2)	(3)	(4)
Focal Citation	-0.509^{***}	-0.346^{***}	-0.561^{***}	-0.562^{***}
	(0.004)	(0.007)	(0.007)	(0.007)
Focal Citation * Exposure (Similarity)			0.095^{***}	
			(0.020)	
Focal Citation * Exposure (Dummy)				0.077^{***}
				(0.016)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
	Filing Lag	Filing Lag	Filing Lag	Filing Lag
Observations	638,142	134,069	265,965	265,965
Adjusted R^2	0.128	0.133	0.117	0.117

*p<0.1; **p<0.05; ***p<0.01

Notes: This specification restricts the sample to citations filed within five years of the focal patent to address concerns about compositional changes in the citing population over longer horizons. Errors are clustered at the matched-set.

Table A.25: Replication of Table 4 with Two-Way Clustering

	Average Distance (in logs)			
	All	High Disclosure	No Disclosure	No Disclosure
	(1)	(2)	(3)	(4)
Focal Citation	-0.529^{***}	-0.386^{***}	-0.624^{***}	-0.624^{***}
	(0.013)	(0.018)	(0.029)	(0.029)
Focal Citation * Exposure (Similarity)			0.136^{***}	
			(0.037)	
Focal Citation * Exposure (Dummy)				0.110^{***}
				(0.029)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
	Filing Lag	Filing Lag	Filing Lag	Filing Lag
Observations	1,136,908	263,794	463,646	463,646
Adjusted R^2	0.133	0.134	0.124	0.124

*p<0.1; **p<0.05; ***p<0.01

Notes: This specification uses two-way clustering at the matched-set and filing-lag levels to account for potential correlation across both dimensions.

Table A.26: Replication of Table 4 with Three-Way Fixed Effects

	Average Distance (in logs)			
	All	High Disclosure	No Disclosure	No Disclosure
	(1)	(2)	(3)	(4)
Focal Citation	-0.529^{***}	-0.386^{***}	-0.624^{***}	-0.624^{***}
	(0.003)	(0.005)	(0.005)	(0.005)
Focal Citation * Exposure (Similarity)			0.136^{***}	
			(0.013)	
Focal Citation * Exposure (Dummy)				0.110^{***}
				(0.011)
Period	2005-2024	2005-2024	2005-2024	2005-2024
Fixed Effects	Matched Set	Matched Set	Matched Set	Matched Set
	Filing Year	Filing Year	Filing Year	Filing Year
	Grant Year	Grant Year	Grant Year	Grant Year
Observations	1,136,908	263,794	463,646	463,646
Adjusted R^2	0.133	0.134	0.124	0.124

*p<0.1; **p<0.05; ***p<0.01

Notes: This specification replaces filing-lag fixed effects with separate cited-year and citing-year fixed effects. Errors are clustered at the matched-set level.