GitPat: A Database Linking Open Source Contributions & Patenting Activity of Organizations

Sergio Petralia

Papers in Evolutionary Economic Geography

24.37

Utrecht University Human Geography and Planning

GitPat: A Database Linking Open Source Contributions & Patenting Activity of Organizations

ABSTRACT

This article outlines a method to link organizations' patenting activities at the United States Patent and Trademark Office (USPTO) with their Open Source Software (OSS) contributions in *GitHub*, the most popular code-hosting service platform. It also provides two ready-to-use databases that are easy to connect to related data sources. The first includes information about all contributions (6,091,653) made to 54 of the most popular OSS projects until June 2024, amounting to over 49 million file changes and more than 3.3 billion line modifications. The second includes information on patents granted until June 2024 (1,719,510) to 1,328 organizations with activity in *GitHub*. This novel data can be used to explore the dynamics and mechanisms driving innovation within modern technological ecosystems, where the lines between proprietary and open-source development are becoming blurry. It offers an opportunity to investigate several unresolved puzzles in the economics of OSS literature, such as disentangling the intrinsic and extrinsic motivations behind individual contributions to OSS, understanding the strategic reasons organizations engage in OSS, and exploring collaboration and geographical concentration mechanisms in the production of digital technologies.

Background & Summary

Open Source Software (OSS) has fundamentally transformed the technological landscape, becoming essential in many sectors. A recent survey by Red Hat found that 82% of Information Technologies (IT) leaders are more likely to select vendors who contribute to the open-source community (https://www.redhat.com/en/resources/state-of-enterprise-open-source- The adoption of OSS goes beyond individual applications, as it is a crucial element of some of the most dynamic sectors, such as cloud computing (https://www.cncf.io/reports/cncf-annual-survey-2023/). The growing importance of collaborative development environments becomes evident by the popularity of code-hosting service platforms like *GitHub*, which has surpassed 100 million developers in 2023 (https://github.blog/2023-01-25-100-million-developers-and-coun

This article describes a method to link organizations' patenting activity at the United States Patent and Trademark Office (USPTO) and *GitHub*, the most popular code-hosting service platform. It also, provides two ready-to-use databases that are easy to connect to related data sources.

GitHub houses the most extensive collection of code repositories, spanning a variety of different programming languages, frameworks, and libraries, across various domains such as web development, machine learning, data analysis, and mobile applications. One of the key features of *GitHub* is its smooth integration with *git*, a distributed version control system that facilitates tracking changes in code, managing different versions of a project, and collaborating effectively with others. Through *git*'s branching and merging capabilities, developers can work on different features or bug fixes simultaneously and merge their changes seamlessly. *GitHub's* public recording of contributions and modifications makes it possible to gain access to the entire development history of any project to study code modification patterns, identify the individuals or teams responsible for specific changes, or understand any aspect of its overall progression.

Given the extensive amount of code repositories hosted on *GitHub*, it is important to note that not all of them hold equal value, as some are plain forks of other projects or personal code collections. In this article I concentrate on projects that offer a digital platform or specific capabilities that could facilitate or enhance innovation. Because of this, only a selected number of projects are considered, those that can be categorized as operating systems, database management frameworks, distributed computing architectures, blockchain technologies, or machine learning libraries, for example. In addition, only projects with a substantial user or contributor base are included. However, the procedure implemented in this article is simple enough such that the resulting data can be easily updated and new OSS projects included, which I plan do regularly. Table 1 shows a detailed list of all OSS projects included (54 in total) along with broad categorization and a brief description of them. The final collection of projects includes a wide variety of domains with a rich longitudinal coverage in terms of their emergence and development over time.

Organizations' activity on *GitHub* is linked to their patent information at the USPTO, which maintains a comprehensive repository on patent applications, granted patents, and trademark registrations. This repository includes detailed information such as the names and geographical locations of inventors and assignees (owners), the technological classifications of inventions, citation records, and regular updates on the legal status of patents, among other details.

The novel data in this article provides a unique opportunity to explore the dynamics and mechanisms driving innovation

within modern technological ecosystems, where the lines between proprietary and open-source development are becoming blurry. It allows for a deeper investigation into several unresolved puzzles in the economics of OSS literature (1; 2; 3; 4). In combination with other sources of data, it can be leveraged to better understand the intrinsic and extrinsic motivations behind individual contributions to OSS by evaluating programmer performance within organizations (5), or developers' propensity to contribute to projects based on social identification or pleasure (6; 7). Additionally, it can shed light on whether exposure to OSS leads to future job offers, equity in commercial open-source companies, or access to venture capital (8).

This data can also be also used to study whether engaging in open-source projects serves as a strategic approach for firms to tap into extensive networks of specialized individuals and for informal knowledge sharing (9; 10; 11; 12). Moreover, it can be used to explore the benefits for individuals and companies who possess complementary assets or produce complementary goods (12; 13; 14; 15; 16; 17; 18; 19). Finally, it can be used to offer insights on the geography of digital technologies (20; 21), on how collaboration works in digital environments, and on the impact of digital transition policies (https://reform-support.ec.europa.eu/what-we-do/digital-transition_en).

Methods

This section describes the procedure implemented to link organizations' contributions on *GitHub* and their patenting activity at the USPTO, as described in Figure 1.

GitHub Data Collection

Contributors' activity for the list of projects in Table 1 were extracted from GitHub after cloning their repositories. For each repository, I extracted detailed information for all contributions (commits) made since their inception and until June 2024, including the commit hash (unique identifier), author email, commit date, and summary measures on the amount of file changes, insertions, and deletions implemented. This was done by leveraging the capabilities of the *git* version control system, after parsing the output of the 'git log' command to capture all necessary details. The extracted data was then transformed and integrated into a structured format.

The resulting, yet preliminary, dataset includes 6,091,653 commits across all 54 OSS projects, amounting to 49+ million file changes and over 3.3 billion line modifications. Table 2 shows the most popular OSS projects along with a concise category description of them. The column 'Commits' provides information on the number of commits in each project while 'First Commit' shows the year of the first commit recorded in the database. Notably, *Chromium* and *Linux* emerge as the most contributed and active projects. However, newer projects like *TensorFlow* have rapidly gained popularity, securing a leading position on this list.

Step 1: Web Domain Extraction

Following the collection of commit information, individual commits are attributed to organizations using the web domain of the contributor's email. To identify and extract web domains from email addresses, the following procedure was implemented:

- The email addresses of commit authors were split using the '@' symbol. For example, the email address *abc@toshiba.co.jp* was split into *abc* and *toshiba.co.jp*.
- The second half of the split email address was further processed to extract the constituent parts of web domains: the root domain and the suffix (e.g., '.com', '.org'). For example, the root domain of *toshiba.co.jp* is *toshiba* and it's suffixes are *.co* and *.jp*.
- The root domain and the first suffix were then concatenated to form the domain. In this example the resulting domain is *toshiba.co*. Domains that correspond to common email providers such as 'gmail', 'protonmail', 'hotmail', 'yahoo', or 'outlook' are discarded in what follows.

Step 2: Identification of Organization Name(s)

The *GitHub* activity of these organizations (identified on the basis of their website domain) are linked to an organization name using information that is available on two main sources: (i) Compustat, and (ii) the 'WhoIs' database.

Compustat (available at: https://wrds-www.wharton.upenn.edu) is a comprehensive database offering detailed information on publicly traded companies worldwide. It provides unique company identifiers along with information on companies' website domains, their legal names, and their patents granted at the USPTO, among other things. Web domains extracted from *GitHub* are then cross-referenced with those listed in Compustat. When an exact match is found, the name of the organization is retrieved.

On the other hand, the 'WhoIs' database is a widely used internet directory that contains information of registered owners of domain names or IP addresses, such as the registrant's administrative and billing details, as well as technical contact information.

These details are provided by registrars during the domain name registration process. This protocol is then used to retrieve registrant's legal name for all web domains extracted from *GitHub*.

Step 3: Retrieval of Patenting Activity

Once organization names are linked to web domains, the following procedure is implemented:

For web domains with a successful match with Compustat, I use their unique identifier to retrieve all patents granted by the USPTO already provided by them. This only includes patents granted from 2011 to 2019. To broaden the scope, I employ an additional matching process. This involves retrieving patents from assignees whose names, as stated in USPTO patent documents, match any of the already linked organization names provided by Compustat. This matching process uses non-disambiguated assignee names to ensure the retrieval of patents that may not have been initially linked due to variations in name formats.

For each organization name linked to a web domain using the 'WhoIs' database, I perform a matching procedure against non-disambiguated assignee names within the USPTO records. When a positive match is discovered, indicating that the same organization is the owner of both the domain and a patent, a link is made such that all patents under the same organization name are attributed to the corresponding web domain.

The next step consists on leveraging the assignee disambiguation made available by the USPTO to retrieve patents that belong to any of these organizations but that were not initially present in the Compustat database or couldn't be matched using non-disambiguated assignee names due to mispellings. Lastly, a series of manual checks have been performed on this data to ensure that the disambiguation provided by USPTO identify accurate matches and that minor inconsistencies or similarities in organization names do not result in false positive or false negative matches. After this procedure, a total of 1,328 unique organizations with both *GitHub* and patenting activity were identified.

Tables 3 and 4 present the top domains in terms of contributions to OSS projects and patenting activity at the USPTO, respectively. Remarkably, *Linux* stands out as the most contributed project across organizations. Several top patentees such as *Intel* and or *Microsoft* are also main contributors of OSS projects like *Linux* and *TypeScript*. Note that in Table 4 web domains are oftentimes grouped, this is because certain organizations like 'Apple Incorporated' have registered under their ownership more than one web domain ('apple.com' and 'webkit.org').

Data Records

The data repository of this article (22) contains two databases and one replication package, for which you can find a detailed description below.

The database called 'OSS.Contributions.csv' contains information about all contributions (i.e. all contributions ever recorded in *GitHub*) to the OSS projects included in this study, which are listed in Table 1. It contains eight columns and 6,091,653 rows. Each row corresponds to a commit in a OSS project while the columns provide the information described in Table 5.

To preserve developer privacy, I don't provide email information of the commiter. Researchers interested in retrieving this data can resort to the replication package, which includes the codes to extract this information.

The second database, called 'PatentData.csv', contains the patent information of the of 1,328 unique organizations with both *GitHub* and patenting activity. This database contains three columns and 1,732,456 rows. Each row corresponds to a patent assigned to a particular we domain or group of web domains. The columns provide the information described in Table 6.

In addition, this repository contains a replication package ('ReplicationPackage.zip') with the code (written in R software language) and the necessary data to replicate the procedure.

Technical Validation

There are several aspects that could compromise the reliability of the matching procedure linking the *GitHub* and patenting activity of organizations. First, errors might occur in assigning organization names to web domains appearing on *GitHub* (Step 2 in Figure 1), leading to incorrect or unassigned organization names for certain web domains. Additionally, during Step 3 in Figure 1, mistakes can arise if the USPTO's disambiguation incorrectly groups different organizations together. Furthermore, at any stage of the procedure, matches might fail due to misspellings in organization names. In the following analysis, I evaluate the presence of false positives (incorrect links) and false negatives (missing links) in the final database.

Assesment of False Positives

To evaluate the accuracy of the database links, I randomly selected 50 entries from the complete set of web domain and organization name combinations (4736 in total). I manually verified whether these assignments were correct. The randomly selected entries are presented in Table 7, where organization names appear as they do in patent documents (in lowercase,

without spaces or special characters). This random sampling approach includes both large patentees, typically well-represented in databases like Compustat, and smaller entities, which are more prone to misassignments. Randomly selecting candidates ensures a comprehensive check across different types of organizations.

After carefully considering this links no obvious mistakes have been found, even for those matchings in which the domain and the organization name differ. For instance, the organization name 'transrol' corresponds to a company based in Chambéry, France, that operates under the name 'Transrol SKF' and is part of the larger SKF Group (skf.com). Similarly, Siebel Systems (siebelsystemsinc) forms part of Oracle (oracle.com) Corporation and Nebbiolo Technologies (nebbiolotechnologiesinc) is owned by TTTech Industrial Automation AG (tttech-industrial.com)

Assesment of False Negatives

To inspect possible missing links in the database I followed a different approach, instead of randomly selecting web domains from the list of all available web domains in the 'OSS.Contributions.csv' database, I purposedly selected the top 50 web domains in terms of commits to any of the projects. This is because it is more likely to find false negatives (organizations that have patenting activity at the USPTO but were not included) among larger organizations. Since it is not obvious how to assign an organization name to these websites I instructed ChatGPT 4 to provide me with a candidate organization name for those domains that are not from an email provider. Table 8 shows the complete list.

As before, no obvious missing links have been found. For instance, note that ChatGPT assigns 'Google' as the organization for all 'chromium.org' contributions. While it is true that the *Chromium* project is mostly maintained by *Google*, not all contributors using a chromium.org email address are employed by *Google*. Additionally, note that although Inktank (inktank.com) is owned by Red Hat (a company that has patents granted by the USPTO), Inktank itself does not have patents and was rightfully excluded, since the acquisition happened in 2014. A similar situation applies to 'freescale.com' and NXP Semiconductors.

Usage Notes

The two databases provided can be connected using the web domain information. This variable is named slightly differently in each dataset to prevent inappropriate merging. In the 'OSS.Contributions.csv' database, each row represents a contribution by an individual, so each contribution (or row) is associated with a single web domain. Conversely, in the 'PatentData.csv' database, each row represents a patent document from an organization, which may be linked to one or more web domains. To prevent row duplication and correctly identify organizations, the web domains are concatenated into a single string, delimited by the character 'l'.

The PatentData.csv database is ready to be merged with any of the datasets available from the USPTO (see https://patentsview.org/download/data-download-tables) using the 'patent_id' variable. This integration allows for the identification of patent owners (assignees), inventors, technological classifications, citations, and the geographical locations of inventors and assignees, as well as access to the full text of each patent document, among other details.

The replication package includes the necessary code to recreate Figure 2, which provides an overview of the patenting activity by the organizations identified through this article's procedure (those who also contribute to *GitHub* projects). Panel (A) shows the share of all patents that are granted to these organizations each year, while Panel (B) shows the technological domains in which these organizations are most active. The percentage of patents granted to organizations that also have commit activity in *GitHub* is approximately 25% for the last two decades. As expected, this share has been increasing along with the predominance of digital technologies. This trend is clearly demonstrated in panel (B), which shows that these organizations predominantly contribute to areas related to computing, information storage, communication, and semiconductor technologies.

The patent data provided in this article can be also easily linked to the historical patenting activity of organizations (23; 24; 25) using patent document identifiers and/or by matching organization names. The same identifier can be used to retrieve high-resolution geolocation data and disambiguated inventor and assignee information (26; 27) or technological classifications can be used to link it to technology-specific measures (28; 29). In addition, *GitHub* repository information can be augmented using ready-to-use procedures to retrieve daily statistics on repositories (https://github.com/DepressionCenter/GitHub-Usage-Stats) or developer activity (30). Note that *GitHub* provides an easy-to-use Access Point Interface (API) to retrieve repository or developer information (https://docs.github.com/en/rest).

Code availability

All procedures implemented in this project were written in R software. I used the following packages: stringr, stringi, data.table, dplyr, plyr, readr, urltools, purrr, scraEP and tm. I provide a simplified version of the original code to facilitate the reproduction of the procedures described in this article, under the name 'ReplicationPackage' (22). The entire code, which is designed to operate on a local computing cluster, is available upon request.

References

1. Lerner, J. & Tirole, J. The open source movement: Key research questions. *Eur. economic review* **45**, 819–826 (2001).

2. Lerner, J. & Tirole, J. Some simple economics of open source. *The journal industrial economics* **50**, 197–234 (2002).

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

4. Rockett, K. Property rights and invention. In *Handbook of the Economics of Innovation*, vol. 1, 315–380 (Elsevier, 2010).

5. Lakhani, K. R. & Wolf, R. G. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Open Source Softw. Proj. (September 2003)* (2003).

6. Bitzer, J., Schrettl, W. & Schröder, P. J. Intrinsic motivation in open source software development. *J. comparative economics* **35**, 160–169 (2007).

7. Bagozzi, R. P. & Dholakia, U. M. Open source software user communities: A study of participation in linux user groups. *Manag. science* **52**, 1099–1115 (2006).

8. Blind, K. *et al.* The impact of open source software and hardware on technological independence, competitiveness and innovation in the eu economy. *Final. Study Report. Eur. Comm. Brussels, doi* **10**, 430161 (2021).

9. Henkel, J. Selective revealing in open innovation processes: The case of embedded linux. *Res. policy* **35**, 953–969 (2006).

10. Dahlander, L. & Magnusson, M. How do firms make use of open source communities? *Long range planning* **41**, 629–649 (2008).

11. West, J. & Gallagher, S. Challenges of open innovation: the paradox of firm investment in open-source software. *R&d Manag.* **36**, 319–331 (2006).

12. Johnson, J. P. Open source software: Private provision of a public good. *J. Econ. & Manag. Strateg.* **11**, 637–662 (2002).

13. Fosfuri, A., Giarratana, M. S. & Luzzi, A. The penguin has entered the building: The commercialization of open source software products. *Organ. science* **19**, 292–305 (2008).

14. Bessen, J. & Maskin, E. Sequential innovation, patents, and imitation. *The RAND J. Econ.* **40**, 611–635 (2009).

15. Polanski, A. Is the general public licence a rational choice? *The J. Ind. Econ.* **55**, 691–714 (2007).

16. Henkel, J. The jukebox mode of innovation-a model of commercial open source development. *Available at SSRN 578142* (2004).

17. Scotchmer, S. Openness, open source, and the veil of ignorance. *Am. Econ. Rev.* **100**, 165–171 (2010).

18. D'Antoni, M. & Rossi, M. A. Appropriability and incentives with complementary innovations. *J. Econ. & Manag. Strateg.* **23**, 103–124 (2014).

19. Tesoriere, A. & Balletta, L. A dynamic model of open source vs proprietary r&d. *Eur. Econ. Rev.* **94**, 221–239 (2017).

20. Wachs, J., Nitecki, M., Schueller, W. & Polleres, A. The geography of open source software: evidence from github. *Technol. Forecast. Soc. Chang.* **176**, 121478 (2022).

21. Wachs, J. Digital traces of brain drain. *EPJ Data Sci.* **12** (2023).

22. Petralia, S. GitPat Dataverse. *Harvard Dataverse* https://doi.org/10.7910/DVN/UQNVHF (2024).

23. Petralia, S., Balland, P.-A. & Rigby, D. L. Unveiling the geography of historical patents in the united states from 1836 to 1975. *Sci. data* **3**, 1–14 (2016).

24. Petralia, S., Balland, P.-A. & Rigby, D. L. Histpat dataset. *Harvard Dataverse* https://doi.org/10.7910/DVN/BPC15W (2016).

25. Petralia, S., Kemeny, T. & Storper, M. Histpat inventor assignee dataset. *Harvard Dataverse* https://doi.org/10.7910/DVN/FQWKGF (2023).

26. Morrison, G., Riccaboni, M. & Pammolli, F. Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Sci. data* **4**, 1–21 (2017).

27. De Rassenfosse, G., Kozak, J. & Seliger, F. Geocoding of worldwide patent data. *Sci. data* **6**, 260 (2019).

28. Petralia, S. Mapping general purpose technologies with patent data. *Res. Policy* **49**, 104013 (2020).

29. Petralia, S. Data from "mapping general purpose technologies with patent data". *Harvard Dataverse* https://doi.org/10.7910/DVN/PQGHKA (2020).

30. Schueller, W., Wachs, J., Servedio, V. D., Thurner, S. & Loreto, V. Evolving collaboration, dependencies, and use in the rust open source software ecosystem. *Sci. Data* **9**, 703 (2022).

Author contributions statement

The author confirms sole responsibility for the following: Conceptualization; Data curation; Methodology; Validation; Visualization; and Writing.

Competing interests

The author declares no competing interests.

Figures & Tables

	Project	Category	Description
1	Ansible	Configuration Management	Tool for automating IT tasks
2	AROS	Operating System	Lightweight, user-friendly operating system
3	Axios	HTTP Library	Library for making HTTP requests
4	Beam	Data Processing	Processes and analyzes large data streams
5	Bitcoin	Blockchain	Digital currency for online transactions
6	Brave	Web Browser	Web browser
7	Cassandra	Data Processing	Highly scalable NoSQL database
8	Ceph	Storage	Distributed storage system
9	Corda	Blockchain	Platform for secure business transactions
10	CouchDB	Data Processing	Database that stores data as documents
11	Chromium	Web Browser	Web browser
12	Dubbo	Microservices	Framework for building microservices
13	ElasticSearch	Search Engine	Engine for searching and analyzing data
14	Ethereum	Blockchain	Blockchain for decentralized applications
15	FastAI	Machine Learning	Library for building AI models
16	FDOS	Operating System	Free operating system compatible with DOS
17	Flink	Data Processing	Framework for processing real-time data
18	Flutter	UI Toolkit	software development framework
19	FreeRTOS	Operating System	Operating system for real-time applications
20	Genode	Operating System	Secure and modular operating system
21	Hadoop	Data Processing	Framework for big data processing
22	Haiku	Operating System	Open-source desktop operating system
23	HelenOS	Operating System	Experimental operating system
24	Hyperledger	Blockchain	Tools for building blockchain applications
25	Illumos	Operating System	Unix-like operating system
26	Istio	Service Mesh	Manages and secures microservices
27	Jenkins	Continuous Integration	Tool for automating software builds
28	Kafka	Messaging	System for managing message streams
29	Kubernetes	Container Orchestration	Orchestrates containers in clusters
30	Laravel	Web Framework	Framework for PHP web applications
31	LightGBM	Machine Learning	Library for gradient boosting in ML
32	Linux	Operating System	Popular open-source operating system
33	Logstash	Log Management	Tool for managing and analyzing logs
34	Mesos	Container Orchestration	Tool for managing computer clusters
35	Mozilla	Web Browser	Web browser and internet suite
36	NetBSD	Operating System	Unix-like operating system
37	NetSurf	Web Browser	Lightweight web browser
38	NodeJS	Runtime Environment	JavaScript runtime for building server-side applications
39	NuttX	Operating System	Real-time operating system for embedded devices
40	OpenBSD	Operating System	Secure, advanced operating system
41	OpenCV	Computer Vision	Library for computer vision tasks
42	OpenFaaS	Serverless	Platform for serverless functions
43	OpenMPI	Parallel Computing	Library for parallel computing
44	OpenStack	Cloud Computing	Manages cloud infrastructure
45	pyTorch	Machine Learning	Library for deep learning
46	React-Native	Web Framework	Framework for building mobile apps
47	React	Web Framework	Library for building user interfaces
48	ReactOS	Operating System	Open-source alternative to Windows
49	SPARK	Data Processing	big data processing framework
50	TensorFlow	Machine Learning	Library for building AI models
51	Terraform	Infrastructure as Code	Tool for defining and provisioning infrastructure
52	Thrift	RPC Framework	Framework for remote procedure calls
53	TypeScript	Programming Language	Superset of JavaScript with static types
54	ZooKeeper	Coordination Service	Service for coordinating distributed systems

Table 1. Full List of Projects

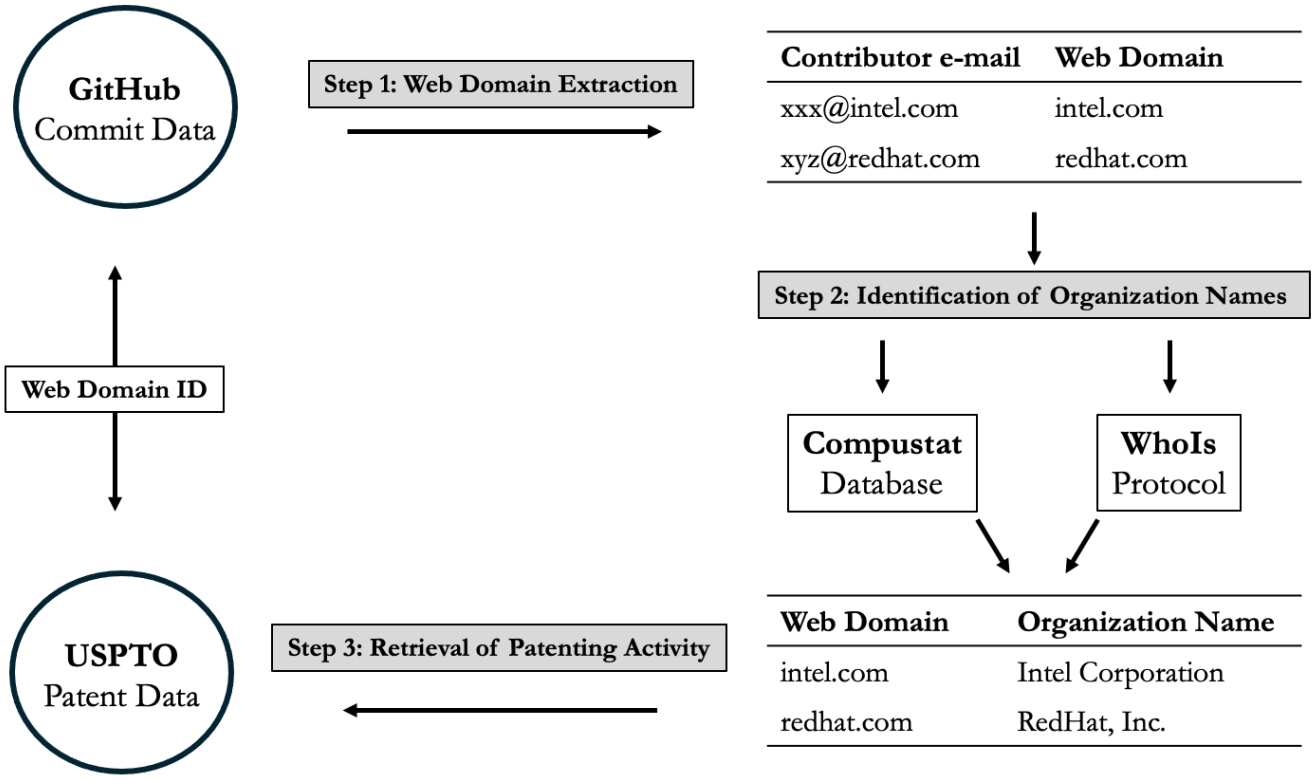

Figure 1. This figure provides an overview of the procedure implemented to link organizations' patenting activity to their OSS contributions in *GitHub*.

	Project	Commits	First Commit	Category
1	Chromium	1,442,750	2001	Web Browser
2	Linux	1,281,048	2005	Operating System
3	Mozilla	922,525	1970	Web Browser
4	NetBSD	312,413	1992	Operating System
5	OpenStack	261,148	2013	Cloud Computing
6	OpenBSD	234,055	1995	Operating System
7	TensorFlow	166,353	2015	Machine Learning
8	Ceph	146,353	2001	Storage
9	Kubernetes	123,627	2014	Container Orchestration
10	ReactOS	85,638	1996	Operating System
11	ElasticSearch	78,239	2010	Search Engine
12	Haiku	64,867	2002	Operating System
13	pyTorch	58,101	2012	Machine Learning
14	Ansible	54,469	2012	Configuration Management
15	NuttX	53,374	2007	Operating System

Table 2. This table presents the top 15 OSS projects in the database along with a concise categorization of them. The column 'First Commit' shows the year of the first commit in the sample, while the column 'Commits' provide information on the number of commits for each project from its inception until June 2024

	Web Domain	Project	Commits	File Changes
1	intel.com	Linux	102,387	241,323
2	redhat.com	Ceph	92,463	377,955
3	redhat.com	Linux	61,267	158,559
4	github.com	Kubernetes	39,409	409,207
5	microsoft.com	TypeScript	24,590	1,355,706
6	elastic.co	ElasticSearch	20,482	287,827
7	huawei.com	Linux	20,087	30,742
8	ti.com	Linux	15,954	32,274
9	redhat.com	OpenStack	14,330	14,330
10	samsung.com	Linux	12,390	25,362
11	nvidia.com	Linux	12,165	24,811
12	mit.edu	Mozilla	11,437	100,860
13	redhat.com	Kubernetes	11,385	92,254
14	opera.com	Chromium	9,726	84,874
15	broadcom.com	Linux	8,827	20,386

Table 3. This table 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 from its inception until June 2024

	Web Domain	Patents
1	iba-group.com |ibm.com	159,298
2	samsung.com	144,761
3	fujitsu.com	59,953
4	toshiba.co	58,337
5	hitachi-solutions.com |hitachi.co |hitachi.com	58,141
6	sony.com |sonymobile.com	56,907
7	ge.com |gehealthcare.com	51,250
8	intel.com	51,190
9	microsoft.com	47,886
10	hp.com |hpe.com |sgi.com	39,661
11	mentor.com |siemens-healthineers.com |siemens.com	38,348
12	micron.com	36,958
13	qualcomm.com |quicinc.com	35,750
14	nec.co |nec.com	35,722
15	apple.com |mac.com |webkit.org	30,414

Table 4. This table shows the top 15 web domains ranked by number of patents at the USPTO

Variable	Description
Project	Name of the OSS project
CommitHash	Commit hash ID
Date	Contribution timestamp, as provided by *GitHub*
Year	Year of commit
FilesChanged	Number of files changed
Insertions	Number of line insertions
Deletions	Number of line deletions
WebDomain	Domain extracted from commiter's email

Table 5

Variable	Description
OrganizationName	Raw assignee name, in lowercase and without spaces or special characters.
patent_id	Patent identifier as provided by USPTO
WebDomains	Web domain(s) associated to the organization

Table 6

	Organization Name	Web Domain(s)	Patents
1	stichtinghetnederlandskankerinstituutantonivanleeuwenhoekziekenhuis	nki.nl	9
2	tycoelectronicsampkorealimited	te.com	1
3	carlzeissmeditecinc	zeiss.com	168
4	universityofkent	kent.ac	19
5	hitachimizusawaeleccoltd	hitachi-solutions.com |hitachi.co |hitachi.com	1
6	honeywellnormalairgarrettholdingslimited	honeywell.com	21
7	sonyelectronicsinc	sony.com |sonymobile.com	2,852
8	lexmarkinternationaltechnologysarl	lexmark.com	6
9	secunetsecuritynetworksaktiengesellschaft	secunet.com	2
10	sonycorportation	sony.com |sonymobile.com	4
11	internatioinalbusinessmachinescorpoation	iba-group.com |ibm.com	1
12	universidadeestadualdecampinas	unicamp.br	6
13	sierrawireless	sierrawireless.com	15
14	sensirionag	sensirion.com	111
15	wistroncorp	wistron.com	492
16	toykoshibauraelectriccoltd	toshiba.co	2
17	generalelectroccompany	ge.com |gehealthcare.com	1
18	volkswagonag	volkswagen.de	11
19	opowerinc	opower.com	42
20	spectralogicinc	spectralogic.com	1
21	babcockhitachikk	hitachi-solutions.com |hitachi.co |hitachi.com	28
22	telefonaktoebolagetlmericsson	ericsson.com	2
23	nuvotontechnologycorporation	nuvoton.com	407
24	templeuniversity	temple.edu	60
25	toshibaengineeringcoltd	toshiba.co	1
26	transrol	skf.com	4
27	kalray	kalrayinc.com	20
28	fabric7systemsinc	fabric7.com	1
29	internationaalbusinessmachinescorporation	iba-group.com |ibm.com	1
30	nebbiolotechnologiesinc	tttech-industrial.com	8
31	netskopeinc	netskope.com	145
32	hitachideviceengineeringcompanyltd	hitachi-solutions.com |hitachi.co |hitachi.com	1
33	hitachiplanttechnologiesltd	hitachi-solutions.com |hitachi.co |hitachi.com	126
34	analogdevicesas	analog.com	1
35	hewlettpackardcomopany	hp.com |hpe.com |sgi.com	1
36	googleinc	google.com |skia.org |tensorflow.org	14,727
37	fujixeorxcoltd	fujixerox.co	1
38	tomtomdevelopmentgermanygmbh	tomtom.com	6
39	trusteesofthelelandstanfordjunioruniversity	stanford.edu	1
40	thetrusteesofthelelandstanfordjunioruniversity	stanford.edu	7
41	sap	sap.com	4
42	lematerieltelephoniquethomsoncsf	gemalto.com |thalesgroup.com	15
43	sonymagnescalecorporation	sony.com |sonymobile.com	1
44	cityuniversity	city.ac	5
45	hitachiulsisystemscoltd	hitachi-solutions.com |hitachi.co |hitachi.com	288
46	internationanlbusinessmachinescorporation	iba-group.com |ibm.com	3
47	koninlijkephillipelectronicsnv	kpn.com	1
48	dialogsemiconductorgmbh	diasemi.com	315
49	fujitsulimted	fujitsu.com	24
50	siebelsystemsinc	oracle.com |sun.com	298

Table 7. This Table shows 50 randomly selected entries from all possible web domain(s) and organization name(s) combinations in the final database. The Column 'Organization Name' displays non-disambiguated organization names as they appear in patent documents (in lowercase, without spaces or special characters). The Column 'Web Domains' shows the web domain(s) that have been linked to the organization name while the column called 'Patents' shows the number of patents in that name-web domain combination.

	Web Domain	Organization Name
1	chromium.org	Google LLC
2	gmail.com	Email Provider
3	netbsd.org	The NetBSD Foundation
4	openbsd.org	The OpenBSD Project
5	gserviceaccount.com	Google LLC
6	openstack.org	OpenStack Foundation
7	kernel.org	The Linux Foundation
8	apache.org	The Apache Software Foundation
9	opendev.org	OpenDev
10	reactos.org	ReactOS Foundation
11	linux-foundation.org	The Linux Foundation
12	linaro.org	Linaro Limited
13	suse.de	SUSE
14	nuttx.org	Nuttx Project
15	gmx.de	Email Provider
16	suse.com	SUSE
17	igalia.com	Igalia
18	davemloft.net	David S. Miller
19	inktank.com	Red Hat (Inktank)
20	genode-labs.com	Genode Labs
21	infradead.org	Infradead
22	linux.org	Linux.org
23	arndb.de	Arnd Bergmann
24	crisal.io	Crisal
25	pengutronix.de	Pengutronix
26	pinc-software.de	Pinc Software
27	glandium.org	Mike Hommey
28	linutronix.de	Linutronix GmbH
29	lst.de	Linux Software Testing
30	itseez.com	Itseez (Intel)
31	outlook.com	Email Provider
32	netsurf-browser.org	NetSurf Browser Project
33	open-mpi.org	Open MPI Project
34	googlemail.com	Email Provider
35	canonical.com	Canonical Ltd.
36	163.com	Email Provider
37	chris-wilson.co	Chris Wilson
38	visionengravers.com	Vision Engravers
39	kohsuke.org	Kohsuke Kawaguchi
40	d-toybox.com	D-ToyBox Project
41	coole-files.de	Coole Files
42	freescale.com	NXP Semiconductors
43	na.email	Email Provider
44	sourceforge.net	SourceForge
45	protonmail.com	Email Provider
46	hotmail.com	Email Provider
47	ideasonboard.com	Ideas on Board
48	bootlin.com	Bootlin
49	cloudbees.com	CloudBees
50	codeaurora.org	Code Aurora Forum

Table 8. This Table shows the top 50 web domains in terms of contributions to OSS projects without patenting activity at the USPTO. The Column 'Web Domain' shows the web domain while the column 'Organization Name' displays the associated organization name that was suggested by ChatGPT 4.

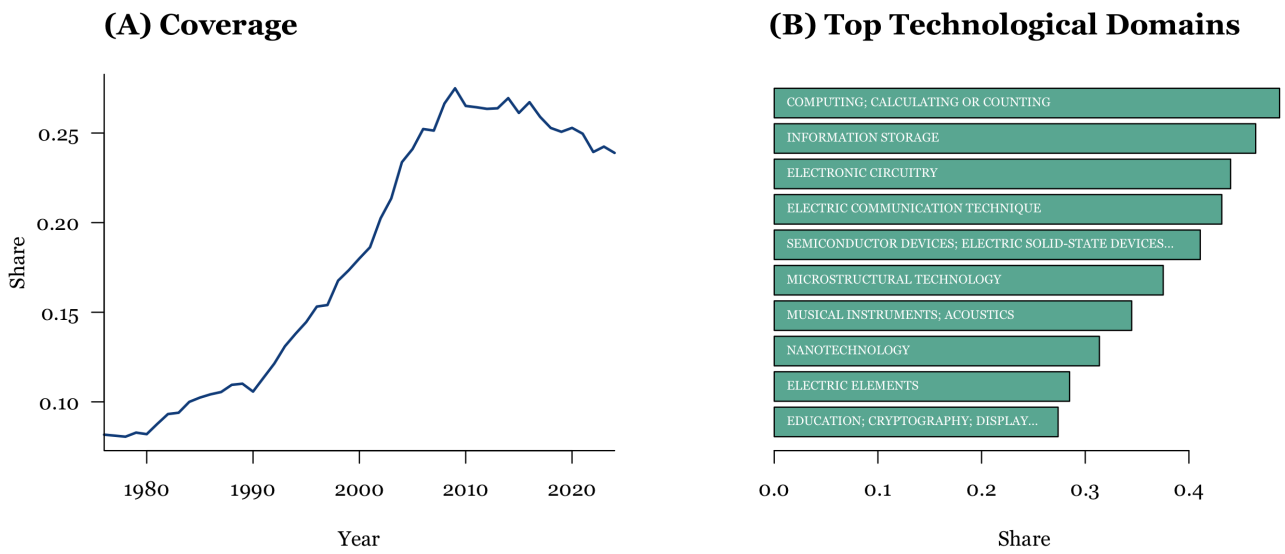

Figure 2. This figure provides an overview of the patenting activity by the organizations identified through this article's procedure (those who also contribute to *GitHub* projects). Panel (A) shows the share of all patents that are granted to these organizations each year, while Panel (B) shows the technological domains in which these organizations are most active.