To replicate Hegde, Herkenhoff, and Zhu (2022), researchers need to access to (1) PATSTAT, (2) PALM Data (via USPTO), (3) Compustat/CRSP, and some supplementary datafiles (e.g., patent-public firm link (obtained from Kogan et al. (2017)) and Google word embedding vectors (via BigQuery)), as noted in the programs.

Main Datasets

us9803.dta: This dataset contains patent applications filed at the USPTO from 1998 to 2003 (inclusive). Observations are application level. It contains pending and abandoned patent applications before and after AIPA. The data is obtained from the USPTO's Patent Application Locating and Monitoring system. This dataset also includes several variables whose construction requires information from PATSTAT such as IPC-based similarity, whether the patent has European or foreign parallel applications.

us9803_public.dta: This the public version of "us9803.dta", where we take out pending and abandoned patent applications (428,690 applications in total). This data can be used to replicate all figures and tables for the event studies (pre-post comparison of U.S. patents) except statistics or analyses related to abandonments (Figure 6 Panel F, Table 1, Table 3, Table C3).

us_ep_twin.dta: This dataset contains US-EP twin patents where the US equivalent is filed at the USPTO from 1998 to 2003 (inclusive) and the EP equivalent is filed at the EPO within 18 months of the US filing date. Observations are at application level and the twins are linked by a unique ID.

us_jp_twin.dta: This dataset contains US-JP twin patents that do not have EP equivalents where the US equivalent is filed at the USPTO from 1998 to 2003 (inclusive) and the JP equivalent is filed at the Japanese Patent Office within 18 months of the US filing date. Observations are at application level and the twins are linked by a unique ID.

rd_int.dta: This dataset contains public firms from 1994 to 2008. The data is downloaded from WRDS compustat/CRSP merged dataset. It can be matched to patents using the patent-firm link from Kogan et al. (2017). Observations are at firm-year level.

ipc7sim_validation_citing_fixed.dta: This dataset contains pairwise ipc7-based cosine similarity between important references, background references, and fake references while keeping the citing patent fixed. Important references refer to X- and Y-citations, as designated by the EPO, which indicate that at least one claim in the citing patent cannot be considered novel or does not involve an inventive step. Background references refer to other types of actual citations. Fake references are constructed by matching the cited patent to another EP patent with the same application quarter, IPC 4-digit code, and grant status.

ipc7sim_validation_cited_fixed.dta: This dataset contains pairwise ipc7-based cosine similarity between important references, background references, and fake references while keeping the cited patent fixed. Important references refer to X- and Y-citations, as designated by the EPO, which indicate that at least one claim in the citing patent cannot be considered novel or does not involve an inventive step. Background references refer to other types of actual citations. Fake references

are constructed by matching the citing patent to another EP patent with the same application quarter, IPC 4-digit code, and grant status.

valid_twin.dta: This dataset contains pairwise IPC7-based cosine similarity and text-based similarity for US-EP twins or fake twins (keeping the US patents fixed and matching EP equivalences to random EP patents with the same application quarter, 4-digit IPC code, and grant status).

Programs

US_EP_DID.do: This Stata program is used to construct all tables and figures that conduct the difference-in-differences using US-EP twins.

US_JP_DID.do: This Stata program is used to construct all tables and figures that conduct the difference-in-differences using US-JP twins and compare the estimates to those from US-EP twins.

US9803_PrePost.do: This Stata program is used to construct all tables and figures that conduct the event studies using US patents filed from 1998 to 2003. When applying this program to the public data we share (us9803_public.dta), one can replicate all figures and tables for the event studies except statistics or analyses related to abandonments (Figure 6 Panel F, Table 3, and Table C3).

RD_Analysis.do: This Stata program is used to construct key variables and conduct the analysis of public firms' R&D investments from 1998 to 2003 in response to AIPA exposure.

EP_Foreign parallel applications.do: This Stata program is used to identify and plot the time trends for US patents with EP or foreign parallel applications (Figure D1). When applying this program to the public data we share (us9803_public.dta) plus PATSTAT data, one can replicate the statistics after excluding pending/abandoned patent applications in Table 1.

US_EP equivalence validation.do: This Stata program is used to compare pairwise similarity between US-EP twins or matched fake twins to validate the equivalence of US-EP twins (Table D1).

Ipc7-sim validation.do: This Stata program is used to compare pairwise similarity between different pairs of citing-cited patents to validate the technology similarity measures based on ipc7 assignments (Table B1).

Exclude software patents.do: This Stata program is used to identify software patents following Graham & Vishnubhakat (2013) and repeat the event studies and difference-in-differences analyses after excluding software patents (Table I1).

Run_all_master_file.do: This Stata program is the master file to run and output all tables and figures in Hegde, Herkenhoff, and Zhu (2022).