Description of Data Files
June 16, 2022

all_patent_data.dta:

This file contains data we frequently use on all issued utility patents (not just financial patents) applied for on or after January 1, 2000 and by December 31, 2018, and awarded by February 28, 1999. Most of this data is taken from the October 8, 2019 Patentsview data release.

The variables are:

- patent_id - the numeric patent number stripped of country and kind codes; can be used to join / merge with other data like financial patent data
- app_date - the application date
- grant_date - the grant date
- primary_cpc - the primary CPC code
- cpc_subclass - the 4-digit (e.g. G06Q) primary CPC subclass (used for calculating mean_cites within cpc_subclass and quarter of grant_date)
- cite_count - the number of forward citations received
- mean_cites - the mean number of cites received by all patents in this data set that have the same CPC subclass and were granted in the same quarter
- weighted_cite - cite_count / mean_cites
- assignee_type - the USPTO type of the first assignee (1 - Unassigned, 2 - US Company or Corporation, 3 - Foreign Company or Corporation, 4 - US Individual, 5 - Foreign Individual, 6 - US Federal Government, 7 - Foreign Government, 8 - US County Government, 9 - US State Government. Note: A "1" appearing before any of these codes signifies part interest)
- kogan_etal_val - the estimated value of the patent in millions of inflation adjusted USD based on stock market response, taken from https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data.
- kelly_etal_val_10 - the estimated value of the patent based on textual analysis with a 10-year forward horizon; computed based on https://www.openicpsr.org/openicpsr/project/119043/version/V1/view.
- kelly_etal_val_5 - the estimated value of the patent based on Kelly et al. (textual analysis with a 5-year forward horizon; computed based on https://www.openicpsr.org/openicpsr/project/119043/version/V1/view.
- csa_code: the numeric CSA code associated with the first inventor's address. The CSAs are taken from an NBER crosswalk associating county FIPS with CSAs, where we use the county FIPS of the first inventor's address as the patent's geographic location.
- csa_title: the description of the CSA region associated with csa_code
- inventor_1_country: the abbreviated country (US, etc.) of the first named inventor
- inventor_1_state: the abbreviated state (CA, etc.) of the first named inventor if in the US
- state_fips: the fips code for the state of the first named inventor if in the US

- vc_num_deals - the aggregate number of VC deals executed in the application year of the patent in the CSA associated with the patents' first inventor
- vc_sum_equity_invested - the aggregate sum of VC money invested (in millions of nominal USD) in the application year of the patent and in the CSA associated with the patents' first inventor
- us_ai_ranking - if the csa_code is one of 6 cities identified as AI hubs, this variable shows the CSA's AI-hub ranking within the US (i.e. the ranking goes from 1-6, and it is not a global ranking); otherwise it is missing. This was not use din the papers

all_patent_data_ICL_ICC_Merged_Marco_V2_20200710.dta:

Same as all_patent_data file, but with the addition of data on the count and mean length of independent claims in the patents at application and issue, as reported in the dataset discussed in Marco, Sarnoff, and deGrazia (2019) and available at https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-claims-research-dataset. For 'number_independent_claim_icc' and 'shortest_independen_claim_icl', which we have for more general number of cases, are generated by Yuan from the Patentsview.

appyear_dummy2:

A simple listing of cells to be used in the decomposition analysis of patent trends.

BS_FintechPatents.dta:

Here we revised the Jinpu Yang's codes as part of the Kelly et al. analyyis, and make it could be running on the end. When the code was running, it should be run from a sequence defined below. I did not change his tf-bdf algorithms, but I did revise some implementation parts. His code documentation (**Documentation for Measure Construction Code.pdf**) should also be carefully examined. Finally, we inserted some of the comments inside each of the script to make it more readable.

• 1terms_count_filtered.py
- Command format: "python terms_count_filtered.py"
- Function: select unigram term count that is in the dictionary

• 2terms_count_parser.py
- Command format: "python terms_count_parser.py"
- Function: parse unigram term count from raw text data and remove all punctuation

•3wc_aggregator.py
- Command format: "python wc_aggregator.py"
- Function: calcuate the tf

• 4dict_aggregator.py
- Command format: "python dict_aggregator.py"

- Function: calculate the backward-df

• 5ComputeSim2019_agg.py
- Command format: “python ComputeSim2019_agg.py fyear1 fyear2”
- Function: calculate patent-level similarity between patents filed in fyear1 and patents filed in fyear2 (fyear1<fyear2, fyear1 and fyear2 are ranging from 1840 to 2019)

• 6ComputeSim2019_pws.py
- Command format: “python ComputeSim2019_pws.py fyear1 fyear2”
- Function: calculate pairwise similarity between patents filed in fyear1 and patents filed in fyear2 (fyear1<fyear2, fyear1 and fyear2 are ranging from 1840 to 2019)

The whole folder for getting the results has a size of 6.6 GB. Yuan uploaded the folder into dropbox \fintech-datasharing\data collection\construct the text measure (BS FS computations)

cite.dta, citeage.dta, citeeconbus.dta, citeeconbustopq.dta, citeit.dta, citepractop3.dta, citetopq.dta, citetop3.dta:

Extracts of the Reliance on Science database (https://zenodo.org/record/4235193#.YIRcU5BKjIV), containing for each patent in all_patent_data, the number of academic citations, the average age of the academic citations (years between the article publication and patent application date), the cites by economics/business/finance journals, the citations by economics in the top quartile by impact factor, the cites by IT journals, the cites bn ”Top 3” practitioner journal (Financial Analysts Journal, Financial Management, and Journal of Portfolio Management), citations by journals in the top quartile by impact factors, and citations by “Top 3” finance academic journals (Journal of Finance, Journal of Financial Economics, and Review of Financial Studies). The impact factors and journal classifications are taken from the Reliance on Science database. Finacademicpreplog.txt in the log/unused folder describes how these variables are constructed.

cited_citing_pairs.dta:

These data were taken from the Rick Townsend database of citations in utility patents issued through December 31, 2019. To be consistent with the rest of the paper, citations in patents after October 8, 2019 were not included. The Townsend data contains citations to non-utility patents, applications, and foreign patent numbers; to the best of our ability these problematic citations were deleted, including deleting cases where the cited patent has a greater number than the citing one, non-numeric strings, and so forth. This data set has the number of the citing patent, the number of the cited patent, and the grant date of the citint patent.

consumer.dta:

We scrutinized the website for the Consumer Financial Protection Bureau and the titles of working papers of the Household Finance Working Group for keywords or bigrams (two-word

phrases) that related to consumer products. (These are listed in Table A-8.) We totaled the number of these keywords or bigrams in the first 100 words of the field labelled "description" or "background" field, the section where these phrases most frequently appeared.

Corporate VC into Finance Firms All Data.xls: The data, which was downloaded from Capital IQ, is described in Appendix D.

corpvcmapping.dta:

The author's mapping of corporate venture capital investment by parent industry, based on Capital IQ industry coding, Compustat, and the author's research.

csa.csv:

A listing of the CSAs used inn the analysis, their states, and the associated codes.

csa_year_data.dta:

This file takes patent level data from financial_patent_data.dta and aggregates it to the CSA-year level, rather than the firm-year level. It also includes Census data at the CSA-year level, including these variables:

- population - total population
- households - number of census households
- median_hh_income - median household income (a weighted a mean of county-level median household incomes in year 2000 only)
- establishments - the number of non-employer establishments in finance/insurance (NAICS 52) according to the economic census
- employees - the number of employees in finance/insurance (NAICS 52) according to the economic census
- adult_population - the total population of those aged 25 or older
- adult_pop_bachelors - the total population of those aged 25 or older with a bachelor's degree or higher level of education

The aggregated patent variables should be self explanatory. They are based on the same simple and weighted patent counts and measures of patent value as described above, but aggregated (sums) at the CSA-year level. The construction of the data series in this field are described in more detail in Appendix G.

fin_addition:

This dataset comprises patents from a group of 278 companies who holds at least one financial patent and presents non-zero or non-missing R&D expenditure data. These details are encapsulated in Tables 12 and Table A-23-26, and further analyses are provided in Appendix H and the related sections in the main body of the paper. The original data was sourced from the

2019 KPSS public patent values, and it was then merged with information about the patent filing year, issued year, and associated citations. We refined this data using financial_patent_data_v3.dta to select only those patents that were financial and had private values and citations. Following this, the refined data was combined with information about each firm's market value, book value, Tobin's Q, and R&D expenditure. This additional information was extracted from the CRSP/Compustat Merged (CCM) database.

finaccites.dta:

Merger between financial_patent_data_v3 and basic information about each academic citation (academic journal ID, publication year, journal impact factor, etc.). Each academic citation is a separate observation. The impact factors and journal classifications are taken from the Reliance on Science database (https://zenodo.org/record/4235193#.YIRcU5BKjIV). Finacademicpreplog.txt in the log/unused folder describes how these variables are constructed.

financial_patent_data_v3.dta:

This file contains selected data on all issued financial utility patents in the same time period as all_patent-data. Because all of the variables reported in all_patent_data.dta are also important for the financial patents, this data set contains all of the variables in that database. It also includes variables showing the fractional share (1 divided by the total number of keyword hits for each patent) attributed to each of our technology classifications, and whether the patent was initially assigned or reassigned to a valid PAE.

These new variables are:

- assignee_1_country – the abbreviated country (US, etc.) of the first named assignee
- accounting - the fractional share in accounting
- investment_banking - the fractional share in investment banking
- commercial_banking - the fractional share in commercial banking
- communications - the fractional share in communications
- payments - the fractional share in payments
- cryptocurrency - the fractional share in cryptocurrency
- currency - the fractional share in currency
- insurance - the fractional share in insurance
- real_estate - the fractional share in real estate
- retail_banking - the fractional share in retail banking
- security - the fractional share in security
- wealth_management - the fractional share in wealth management
- pae_initial_assignment - a binary variable indicating that the first assignee matched to a valid PAE name
- assignee_1_pae_name - the name of the first assignee if pae_initial_assignment == 1
- pae_reassignment - a binary variable indicating whether the patent was reassigned to a valid PAE *at any point in time after Jan. 1, 2000*

- reassignment_pae_name - the name of the ***first*** valid PAE (in time) to which the patent was reassigned, if pae_reassignment == 1
- pae_reassignment_date - the execution date of the ***first*** reassignment if pae_reassignment == 1

The dataset also includes a host of financial information about the first assignee taken from Capital IQ and based on the patent application year unless otherwise noted (some Capital IQ variables report prior year data, in which case we have the year before the application year). These variables are (variables explained in further detail where not obvious):

- capiq_id - the Capital IQ identifier for the first assignee
- company_name - the name of the first assignee from Capital IQ
- revenue
- ebitda
- rd_expense - research and development expenses
- advertising expense
- net_income
- cash - cash in the year prior to the application year
- long_term_debt - long term debt in the year prior to the application year
- short_term_debt - short term debt in the year prior to the application year
- short_term_investments - short term investments in the year prior to the application year
- shareholder_equity - shareholder equity in the year prior to the application year
- market_cap - market capitalization as of the end (12/31) of the application year
- employment
- year_founded
- age_of_firm
- primary_industry - industry name for the first assignee's primary industry
- primary_industry_code - full 8-digit GICS code for the first assignee's primary industry
- gics_sector_code - 2-digit GICS sector code for the first assignee's primary industry
- gics_group_code - 4-digit GICS sector code for the first assignee's primary industry
- gics_industry_code - 6-digit GICS industry code for the first assignee's primary industry
- gics_sector - character variable with sector name corresponding gics_sector_code
- gics_industry_financials - character variable with industry name corresponding to gics_industry code only if gics_industry_code == 40 (for financials)
- revenue_group - a character vector indicating whether the firm's revenue in the application year was less than $100 million ("Small"), more than $100 million but less than $10 billion ("Medium"), or greater than or equal to $10 billion ("Large"). If revenue is missing, this will also be missing.
- public - a binary variable indicating whether the firm had positive, non-missing market capitalization at the end (12/31) of the application year (1); this variable will be equal to 0 in years for which the firm did patent but did not have a market capitalization at the end (12/31) of the application year; it will be missing in all other years (including years in which the firm did not patent)

- global_sifi - a binary variable indicating whether first assignee is a global SIFI
- vc_backed - a binary variable indicating whether the firm received VC-funding at any point after January 1, 2000

finvcmapping.dta:

The authors' mapping of financial venture capital investment by parent industry, based on Capital IQ industry coding, Compustat, and the authors' research. Non-strategic investors are dropped from the analysis.

FS_Fintech_Patents.dta:

See BS_Fintech_Patents.dta above.

FundPatents.dta:

Supplemental identification in patent abstracts with references to keywords associated with active and passive funds (two categories not originally broken out in the coding, but needed to ensure mapping with the BEA 405-industry scheme).

Gross Output Final.xlsx:

Simplified version of gross output file (https://apps.bea.gov/iTable/index_industry_gdpIndy.cfm), whose construction is described in Appendix F.

JournalID.xls:

Renamed and reformatted version of journalidname file and folder compiled by Reliance on Science: https://zenodo.org/record/4235193#.YIRJeZBKjIV.

OldPatentsFinal:

The dataset from Lerner (2002), with the addition of the weighted_cite, kogan_etal_val, and kelly_etal_val_5variables, constructed as noted in the description of all_patents_data file. The assignment of patentee type differs slightly from Lerner (2002), as this classification is now based on USPTO reporting in PatentsView, to be consistent with all_patents_data. The 2002 paper classified patents based on the author's own research. In particular, a small number of patents that were assigned to holding companies associated with a single inventor were classified in that paper as being individual patents, but by the USPTO (and Patentsview) as corporate ones.

patent_vc.dta:

We identify a patent as venture backed if the firm it was assigned to was financed by a VC and its application date is between the assignee's first and last venture round dates. We use Refinitiv VentureXpert data following the approach in Akcigit et al. (2020).

process.dta:

Based on our classification of patent types, we assumed that all communications and security patents are unambiguously process-related ones. For the remaining patents, we divide them into process and product ones following the methodology of Banholzer et al (2019), which focused on the presence of process patent-related keywords in independent claims. We took the most conservative of their measures, which measured the share of independent claims that are process-related based on the initial two keywords.

regulation-related datasets:

The input datasets (`state_level_technology_science_index.dta`, `state_regdata_industry_2-1.csv`, `state_regdata_restrictions_2-1.csv`, `fipscounty_csa_crosswalk.dta`) serve as the foundation for the examination of regulations at the state and CSA levels. These analyses include the study of the effects of geographical shifts and changes in the technology landscape on financial patenting. The summarized results can be found in Table 9-11, Table A-15, and Figures A-8, and A-9. State RegData was sourced from the QuantGov platform while the technology index data is from STSI.

The intermediate files derived from input datasets are `state_regdata_industry_restriction_finance_only_collapse.dta` and `state_indtype_apptype_appyear_with_techindex_cells.dta` for Table 10. Output files corresponding to the benchmark analysis in Table 9 are contained in `00_19_state_level_patent_cell_data_new_regdata.dta`. Meanwhile, the benchmarks established in Table 11 can be found in `finregswitch_csastate_pairs_company_level.dta`.

social_return.dta:

This dataset has been put together in order to investigate both the private and societal benefits of financial patents, as displayed in Figure 7 and Figure A-12-13. To begin with, we utilized the `all_patent_new.dta` dataset, which encapsulates the weighted citation accounting for all inventors. This dataset aids in establishing the denominator in the financial/non-financial ratio. Following that, the dataset was merged with `financial_patent_data_v3.dta`, and supplementary datasets (`public_patval_permco_2019.dta`, `ccm_full.dta`) featuring patent values and R&D expenditure from `fin_addition.dta`, to compute separately the private and societal benefits of financial and non-financial innovation.

Moreover, the St. Louis Fed's API was employed to bring in GDP, TFP, and investment data pertinent to the financial sector. This was paired with the finance fraction to assess the benefits and costs. The `social_return.dta` dataset eventually preserves different ways to calculate the benefits and costs as discussed in the paper, as well as the ultimate ρ.

software.dta:

We followed the methodology employed by Chattergoon and Kerr (2020), which in turn is based on Bessen and Hunt (2007), and again draws primarily on key words in the description field.

state.csv: Mapping of states to abbreviations and to census regions, taken from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

university_patent.dta:

A listing of patents in the all_patent_data file associated with a university. We compiled all patents with an assignees containing the word "university," as well as those on the various annual lists of the most active academic patentees compiled by the Association of University Technology Managers (which allowed us to capture entities as the Massachusetts Institute of Technology and the Wisconsin Alumni Research Foundation).