Reproducing the Analyses and Figures

This README explains how to reproduce all data processing, analyses, and figures for the paper **"Breakthrough Odds Are Double in Emerging Research Lines"** by Keyvan Vakili (London Business School), Michael Bikard (INSEAD), and Florenta Teodoridis (University of Southern California). It is designed so a technically proficient reader can retrace the full pipeline end-to-end using the provided code and public datasets.

At a glance

- Raw data: Microsoft Academic Graph (MAG, 2021-12 snapshot), PatentsView, and Nobel-related papers (Harvard Dataverse).
- Compute: Azure Synapse Analytics (Dedicated SQL pool) with PolyBase external tables reading from Azure Blob Storage.
- Post-processing: Stata (reference implementation). Other statistical software may be used. - Visualization: Tableau Desktop **2024.3**.
- Outputs: **Tableau-ready CSV files (provided)** for each figure, plus intermediate Parquet exports from Synapse.1) Obtain the data

---

1.1 Microsoft Academic Graph (MAG)

- Use the **2021-12 snapshot** (the final public snapshot). The scientific publication data were obtained from the Microsoft Academic Graph (MAG; 202112 snapshot). Microsoft no longer maintains or hosts MAG; the raw files used in this study are available from the authors on request. OpenAlex provides free access to a dataset covering the majority of MAG, but its data structure and files differ and may require conversion to follow our replication guideline.
- Place the following MAG text files into your Azure Blob container with the expected relative paths:
    - `/mag/Papers.txt`
    - `/mag/PaperReferences.txt`
    - `/mag/PaperAuthorAffiliations.txt`
    - `/advanced/PaperFieldsOfStudy.txt`
    - `/advanced/FieldsOfStudy.txt`
    - `/advanced/FieldOfStudyChildren.txt`
    - `/advanced/FieldsOfStudyToTopField.txt`

Background: The original processing codebase used **U-SQL** on Azure Data Lake Analytics, which has since been discontinued. We converted the U-SQL logic to **T-SQL** for **Azure Synapse Analytics**; the scripts in `/scripts/ Converted T-SQL`

`scripts/` implement the one-for-one transformations needed to build all derived variables.

1.2 Nobel-related papers (Dataverse)

- Download the dataset of publication records for Nobel laureates from Harvard Dataverse: `doi:10.7910/DVN/6NJ5RN`.
- Save the relevant files under `/data/external/nobel/` and, if needed, mirror them to your Blob container for server-side joins.

1.3 Patents (PatentsView)

Download current bulk tables from **PatentsView** and copy to your Blob container as tab-delimited files with these exact names (expected by our SQL): - `/application.tsv` - `/patent.tsv` - `/cpc_current.tsv` - `/uspatentcitation.tsv`

---

2) Stage data in Azure Blob & generate a SAS token

You'll need an **Azure Storage account (Blob)** and an **Azure Synapse workspace** with a **Dedicated SQL pool** (formerly SQL DW).

1. Upload all files from Section 1 to an Azure **Blob container** (e.g., `magdata`).
2. In the Azure Portal, generate a **Shared Access Signature (SAS)** for the container with at least **Read & List** permissions (and **Create/Write** if you plan to export from Synapse to the same container). Copy the **query string** form of the token beginning with `?`.
3. Keep handy:
    - `<storage-account-name>` (e.g., `myaccount`)
    - `<container-name>` (e.g., `magdata`)
    - `<sas-token>` (the `?sv=...&sig=...` string)

---

3) Run the Synapse T-SQL pipelines (server-side processing)

Open Synapse Studio (or SSMS) connected to the **Dedicated SQL pool**. Execute the two scripts in `/sql` in the order below, after adjusting placeholders.

3.1 Publications pipeline

- Open **`/sql/TSQL script for processing publications data.sql`**.
- Set the container and account placeholders:
    - `LOCATION = 'wasbs://<container-name>@<storage-account-name>.blob.core.windows.net'`.
    - `SECRET = '<sas-token>'` (include the leading `?`).

- Run the script. It will:
    1) Create external file formats for TSV; 2) Create external tables over MAG; 3) Build typed **internal** tables/views hashed for performant joins; 4) Construct topic-pair frequencies by year; 5) Compute **cohort (lag-1)** and **historical (lag-4 window)** frequency indices at the **paper** level; 6) Build forward-citation windows (3,5,10-year) and **top-5%** cohort flags; 7) Materialize **`PaperDataMerged_int`** and export **`PaperDataMerged`** to **Parquet** under your blob **`/Output/run1/PaperDataMerged/`**.

Result: - **Intermediate Parquet**: `/Output/run1/PaperDataMerged/` (Parquet files)

3.2 Patents pipeline

- Open **`/sql/Converted T-SQL scripts/TSQL script for processing patents data.sql`**.
- Update the same placeholders for container/account/SAS.
- Run the script. It will:
    1) Create external tables over PatentsView TSVs; 2) Build typed **internal** tables; 3) Compute forward citations by application-year window and **top-5%** cohort flags; 4) Construct CPC-pair frequencies and the lagged cumulative counts; 5) Materialize **`PatentsDataMerged_int`** and export **`PatentsDataMerged`** to **Parquet** under **`/Output/patents/run1/PatentsDataMerged/`**.

Result: - **Intermediate Parquet**: `/Output/patents/run1/PatentsDataMerged/`

Why Synapse? The datasets are large and the joins are heavy. The scripts assume hash-distributed internal tables and columnstore indexes to keep performance predictable.

---

4) Define the core measures & quadrants (used across figures)

Let the **cohort frequency index** be `MedianRefPairFreqL1` and the **historical frequency index** be `MedianRefPairCumFreqL4`.

- **Interpretation:** Higher values indicate **lower divergence**; lower values indicate **higher divergence**.
- **Sample-wide medians:** Compute `median_cohort` and `median_historical` over all papers (or patents) within the sample window.

Then classify each work into one of four quadrants:

Quadrant	Divergence from prior research	Divergence from mainstream	Rule (using indices)
Uncharted	High	High	`cohort <= median_cohort` **and** `historical <= median_historical`
Established	Low	Low	`cohort > median_cohort` **and** `historical > median_historical`
Emerging	High	Low	`cohort > median_cohort` **and** `historical <= median_historical`
Classic	Low	High	`cohort <= median_cohort` **and** `historical > median_historical`

Replace `PaperDataMerged` with `PatentsDataMerged` to replicate the patent-side classifications.

Breakthrough ("hit") definitions used in the paper:

- **Papers:** top 5% in forward citations within publication-year cohorts (we report multiple windows; the main text uses the **5-year** window for primary figures).
- **Patents:** top 5% in forward citations within application-year cohorts.
- **Nobel-related papers:** identified using the Dataverse list and cross-referenced to MAG.

---

5) Local processing (normalization, margins, and summary stats)

Our reference implementation uses **Stata** for two reasons: (i) normalizing a few variables prior to Tableau ingestion, and (ii) computing margins/adjusted rates for Figure 5. You may use other statistical software, but we provide no example code here.

5.1 Stata (reference)

1. Ensure `PaperDataMerged` and `PatentsDataMerged` Parquet exports are available locally (or accessible via cloud-mounted paths). Convert to **CSV** if needed using your preferred tool (e.g., `arrow`, `pandas`, or Azure Storage Explorer).
2. **Import the CSVs into Stata manually** (filenames are user-defined and therefore not hard-coded in the do-file):
    - From the menu: **File → Import → Text data (delimited)**, or
    - From the command line: `import delimited using "<your_papers_export.csv>", clear` (and later with your patents export).
3. Open **`/scripts/Stata Do File/Analysis.do`**, set the working directory at the top, and **run the relevant section twice**:

- **Once for papers** (with the papers CSV currently in memory), and
- **Once for patents** (after importing the patents CSV into memory).

The do-file is organized into clearly annotated blocks for papers and patents. It assumes the dataset in memory contains at least the variables listed in Section 3 results (e.g., `MedianRefPairFreqL1`, `MedianRefPairCumFreqL4`, `Top5pct_3yr`, author/field identifiers for papers; analogous CPC/forward-citation fields for patents).

4. The do-file will:
   - Compute medians and classify quadrants (Section 4 rules),
   - Normalize the variables required prior to Tableau ingestion (where applicable),
   - Produce **summary statistics** for Figure 4, and
   - Estimate the model and compute **margins** for Figure 5.
5. **Exports:** The do-file **does not** write CSVs by default. Readers may export or save outputs as needed (e.g., `export delimited`), or use the **provided reference CSV files** that were imported into Tableau for the submission.

Stata version: Tested with Stata 19.5. Make sure your Stata has `margins` and related features.

6) Where to get the final Tableau-ready files

If you prefer to skip local processing, we also provide the finalized, figure-specific **Tableau-ready CSV files** at **`/datasets/`**

These files and any accompanying Tableau workbook let you **inspect exactly how the figures are created** from the exported data—we do not include step-by-step Tableau build instructions in this README.

---

7) Reproducibility notes & expected outputs

- The Synapse scripts are **idempotent**: they drop and recreate external tables/views/materialized tables as needed.
- The **publication** pipeline exports `PaperDataMerged` Parquet; the **patent** pipeline exports `PatentsDataMerged` Parquet. Each folder will contain multiple partition files.
- Local Stata steps compute classifications, summary stats, and margins; **export CSVs yourself if desired**. We also provide **reference CSVs** used in the paper for direct Tableau ingestion.
- You should be able to match (within rounding/plotting tolerances) all reported numbers in the main text and SI.

---

8) Citations & acknowledgements

- **MAG**: Microsoft Academic Graph, 2021-12 snapshot.
- **PatentsView**: U.S. patent data
- **Nobel-related data**: Harvard Dataverse DOI `10.7910/DVN/6NJ5RN`.
- Original U-SQL codebase was ported to T-SQL for Synapse by the authors

---

9) Contact

For questions about data access or reproductions, please contact the corresponding authors listed in the manuscript.

Last updated: 21 Aug 2025