Measuring Technological Innovation over the Long Run (2021)

Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy

Presented by Jieming Zhang, Feb 06 2023

Purpose

- Construct and measure indices of technological progress over the past 2 centuries
- Aggregate and sectors

Existing Method

Patent citations Problems

- Data is incomplete, missing data prior to 1947
- Citations take discrete values
- Rely on subjective discretion by patent examiner (selective bias)

Solution

- analyze texts of patent documents
- range of data availability extended (1840-2010)

Method

Step 1

- NLP to create links of invention between its former and subsequent inventions
- Construct textual similarities to quantify commonality of each pair of patents.
- Identify importance of patents by novelty and influence

Method

Step 2

- Create time series indices to measure intensity of breakthrough innovations at aggregate and sector levels by counting number of most important patents
- The aggregate innovation index uncovers three historical technology breakthroughs
  1. Second Industrial Revolution
  2. 1920s-1930s Electricity and Petroleum
  3. post-1980s Computer Science

Measuring similarity

Data source

- US Patent and Trademark Office
- Google's patent search engine
- 9 million patent texts from 1840-2010

Text similarity

- Weigh words importance by TFIDF

$$TFIDF_{pw} = TF_{pw} \times IDF_w \tag{1}$$

Term Frequency

TF_{pw}
measures how many times term w appears in patent p adjust for length
IDF – Inverse Document Frequency

- measures informativeness of term w by underweighting common words.
- a high IDF indicates term w is informative but not common in other patent documents

TFIDF

Importance of word w given patent p

- High TFIDF means high frequency of term w in patent p but low frequency in other documents.

Limitation

- Limit the influence of breakthrough invention: Cited too much by future documents
- Result in lower IDF

Modification

Term Frequency Backward-Inverse Document Frequency (TFBIDF)

- Limit citation to the earliest time between pair of patents
- Normalize TFBIDF to unit lenght
- Compute cosine similarity ρ from [0,1]

Correlation of Citation and Similarity

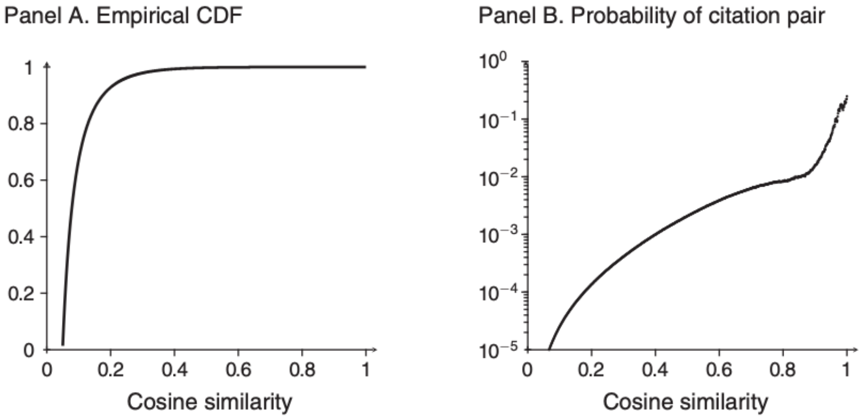


Figure 1. Pairwise Similarity and Citation Linkages

Novelty and Importance

- use similarity ρ to compute backward similarity BS
- the lower the BS the more distinct the patent is from existing patents given a time range of τ
- use ρ to compute forward similarity FS the higher the FS the more influential the patent is to following patents given a time range of τ
- importance indicator combines patent's novelty and influence
  $q_j^\tau = \frac{FS_j^\tau}{BS_j^\tau}$.

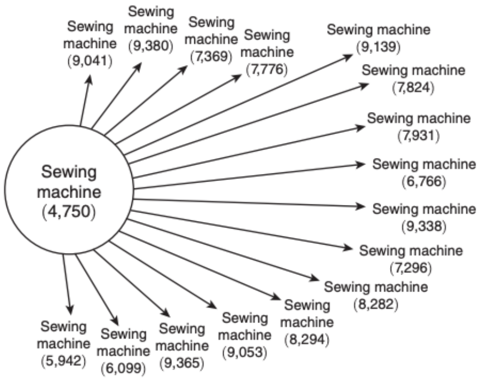
Validity Confirmation

1. Manually Identify a set of technological breakthroughs and compare the significance of indicators
2. Compare indicator to citation numbers after data is available (post-1947)
3. Use indicators to predict future citation numbers
4. Relate citations to private values. Consistent with Kogan et al. (2017)

Comparing technological breakthrough

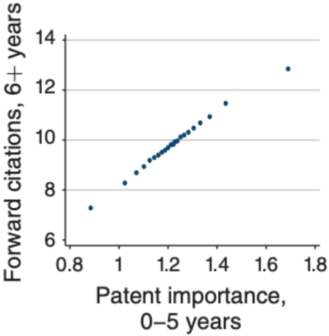
The first patent of sewing machine by Elias Howe Jr.

Panel A

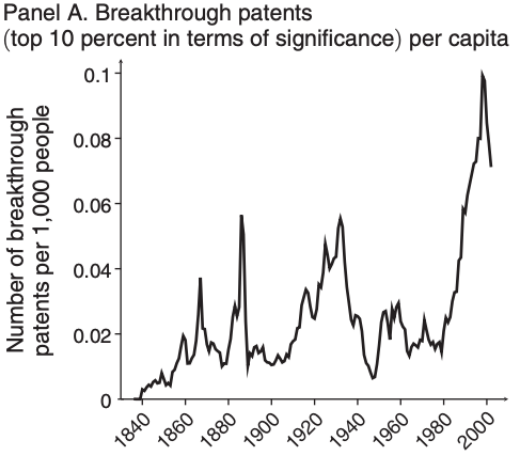


Patent importance and future citations

The first patent of sewing machine by Elias Howe Jr.



Time series of tech breakthroughs per capita



consistent with historical evidence