Duplication in Innovation Efforts over Time: Natural Language Processing Techniques and US Patent Data

Research Proposal — Konstantin Poensgen

November 2022

1 Gap in the Literature

Various empirical studies have documented a decline in research productivity on the aggregate and individual firm level over time (Bloom et al., 2020; Gordon, 2016; Cowen, 2011; Griliches, 1994). This research proposal concerns one potential contributor to this trend: **Has duplication in innovation efforts increased over time?**

Previous work has analyzed the novelty, similarity and impact of innovations using patents as an indicator. The proposed study seeks to add to this literature by asking about duplication across innovation activities. Whereas key metrics of prior research took the view of individual patents, the proposed study seeks to focus on related patents and rejected patent applications holistically, i.e. aggregate from the individual level. Using data derived from machine learning techniques, the proposed study specifically adds to recent work applying text-mining techniques to US patent data (Arts et al., 2020; Hain et al., 2020; Feng, 2020; Ashtor, 2019; Arts et al., 2018).[1]

2 Importance

Innovation is central to generate technological progress and to promote economic growth. Understanding the causes of falling research productivity is thus an important macroeconomic and policy-relevant issue. Increases in duplicated innovation efforts could be one factor contributing to the observed fall in research productivity and has been incorporated in some endogenous growth models (Jones, 1995; Zeira, 2011). As pointed out by Bloom et al. (2020), "doubling the number of researchers may less than double the production of new ideas because of duplication or because of some other source of diminishing returns" (p.1115). Further, identifying covariates of research duplication such as market structure or industry type could provide insights for competition and innovation policy.

3 Research Strategy

3.1 Data

In a first step, the proposed study could utilize data provided by Arts et al. (2020) who use natural language processing tools to extract keywords from granted US utility patents from the USPTO, a patent claims research dataset from Marco et al. (2016), and PATSTAT. Arts et al. use this data to propose a range of text-based metrics to study patent novelty and provide their full data and code open access. The data is available under:

> Arts, S., Hou, J., and Gomez, J.C. (2020). Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures, *Research Policy*.

1 For a recent review see: Antons et al. (2020): The application of text mining methods in innovation research: current state, evolution patterns, and development priorities, Special Issue: Innovation Management Research Methods, *R&D Management*, 50(3), pp. 329-351.

The data provided includes "for each patent a list of processed and cleaned keywords extracted from the title, abstract, and claims. [The] data can be used to measure and map the similarity between patents, inventors, firms, or geographical regions in technology space" (Arts et al., 2020, p.12).

In a second step, a main exercise of the proposed study would be to extend the work by Arts et al. to include rejected patent applications, for instance retrieved from the USPTO Patent Examination Research Dataset (Public PAIR). This would be required to better capture duplication in innovation activity as rejected patent applications could in part embody ideas already existing and patented, thus informing about research duplication. Ideally, the code provided by Arts et al. could be used as a close reference point for this purpose.

3.2 Empirical Strategy

The main empirical assessment would follow a two-step procedure:

1. Similar in spirit to recent work, use patent text keywords to measure similarity across patents in a given time window (e.g., Arts et al., 2018). Similarity measures could be the Jaccard index, cosine similarity, and Mahalanobis distance.
2. Compute measures of duplication for different levels of aggregation such as the North American Industry Classification System (NAICS) over time. Potential candidates are tractable statistics such as weighted averages, median or variance of the similarity measures.

Interesting tests to perform include simple regressions to identify:

- The change in duplication measures over time;
- Covariates within unit of aggregation over time and across units;
- Duplications in rejected applications versus granted patents.

Important robustness checks include:

- Compare text-mining based metrics to traditional measures such as patent classification (e.g. Jaccard index using subclasses) and citation information (e.g. forward citation);
- Time-invariance of the relationship between granted patents, rejected applications and duplication to extrapolate from granted patents to applications for early periods when patent application data is not available.

4 Extensions

- Identify factors underlying the observed duplication trends. For example, one can correlate duplication to market characteristics such as the degree of concentration. This can shed light on whether the decline in productivity is intrinsic or due to the market environment, which in turn can inform policies helping to reverse the decline in innovation productivity.
- Test predictions from Zeira (2011) whether duplications are particularly concentrated in "low cost innovations, while some of the high-cost innovations are under-researched" (p. 137).
- Perform statistical tests to identify structural breaks in trends of research duplication (e.g., after "major" inventions; entry/exit of key market player; after introduction of R&D stimulating policies such as R&D tax credits);
- Expanding the keyword and n-gram NLP techniques developed by Arts et al. (2020) to topic modelling and word embedding.

Bibliography

Antons, D., Grünwald, E., Cichy, P., and Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities, Special issue: Innovation management research methods. *R&D Management*, 50(3):329–351.

Arts, S., Cassiman, and Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Arts, S., Hou, J., and Gomes, J. C. (2020). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*.

Ashtor, J. H. (2019). Investigating cohort similarity as an ex ante alternative to patent forward citations. *Journal of Empirical Legal Studies*, 16:848–880.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). The role of prices relative to supplemental benefits and service quality in health plan choice. *American Economic Review,*, 110(4):1104–44.

Cowen, T. (2011). *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better*. Dutton, New York.

Feng, S. (2020). The proximity of ideas: An analysis of patent text using machine learning. *PLoS Online*, 15(7):e0234880.

Gordon, R. J. (2016). *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*. Princeton University Press, Princeton, NJ.

Griliches, Z. (1994). Productivity, R&D and the data constraint. *American Economic Review*, 84(1):1–23.

Hain, D. S., Jurowetzki, R., Konda, P., and Oehler, L. (2020). From catching up to industrial leadership: Towards an integrated market-technology perspective. an application of semantic patent-to-patent similarity in the wind and EV sector. industrial and corporate change. *Industrial and Corporate Change*.

Jones, C. I. (1995). R&D-based models of economic growth. *Journal of Political Economy*, 103(4):759–84.

Marco, A., Sarnoff, J., and de Grazia, C. (2016). Patent claims and patent scope. *USPTO Economic Working Paper*.

Zeira, J. (2011). Innovations, patent races and endogenous growth. *Journal of Economic Growth*, 16(2):135–156.