Review of: "Strategic Citations in Patents: Analysis Using Machine Learning"

Per Botolf Maurseth[1]

1 Norwegian School of Management BI

Potential competing interests: No potential competing interests to declare.

Introduction

Sijie Feng's paper proposes to use proximity in ideas in patents based on machine learning as an alternative to patent citations to measure relatedness of patents. The idea is to use machine reading of texts in patent documents to find similarities between pairs of patents. Patent documents contain fairly detailed descriptions of the invention in question. Similarity in the description of two patents can be used to determine the degree to which two patents are similar and therefore related to each other. Feng proposes such machine learning similarity indexes as an alternative to patent citations. As will be become clear, I think the proposed similarity index cannot substitute for patent citations, but probably complement measures based on patent citations.

In this short review I first discuss the importance of knowledge spillovers in growth economics. Thereafter, I discuss patent citations and how patent citations have been used in innovation research. I will then summarize Feng's paper and arguments before a short discussion about machine learning similarity indexes versus patent citations.

Technology spillovers and economic growth

Patent citations have become increasingly used as indicators of some relationship between patents. This is important. Both new and older growth theory in economics hypothesize that knowledge flows, between persons, firms, industries, regions and countries, are of crucial importance for our understanding economic growth. Nobel-laurate Trygve Haavelmo (1954) wrote that:

Third, there is the important 'trade' in knowledge, ideas and ideologies. These items have – at least very often – the peculiar and fortunate property that they can be sold and still kept. Where it not for the difficulties of quantitative measurement, one might well hold that 'trade' of this kind has meant more in the process of evolution than all the material goods that have crossed regional borders.

Haavelmo notes the public good characteristic of knowledge, its importance and the problems of measurement. Importantly, pure knowledge spillovers, i.e. diffusion of knowledge that is not subject to market transactions, is believed to be a source of endogenous economic growth. Such spillovers may explain increasing returns at the macro level even when there are decreasing returns at the micro level (Romer, 1986).

Patent citations

Patent citations have been used as one measure of knowledge flows. Patent documents contain references, citations, to existing patents. The citations are added to indicate some link of relevance between the cited and the citing patent. Citations are added both by the patent applicant as well as the patent examiner at the patent offices. The patent citations have the legal purpose of limiting the scope of patent protection.

Patent citations are potential indicators of knowledge spillovers if the knowledge inherent in a cited patent in some way constitutes the knowledge basis used to develop the citing patent. Therefore, patent citations are potentially very useful in empirical studies of the cumulative aspects of knowledge production and as a measure of knowledge diffusion.

Knowledge spillovers as such denote the theoretically distinct and clear concept of knowledge externalities. Useful theories are necessarily simplifications of reality. The interpretation of patent citations as knowledge spillovers has to be qualified. It is well known that practices of adding patent citations vary. Different patent offices, patent examiners and patent attorneys tend to add citations differently. There are also different practices between countries. Also patent applicants add patent citations. Feng argue that applicants add citations strategically.

As described above, patent citations are often added by the patent office and not by the patent applicants. The inventor of a patented innovation may therefore not have been aware of the cited patent. In this case there is no direct spillover link between the patents. Still, there is a citation link that indicates relevance between the two patents. Relevance of knowledge alone should be expected to trigger spillovers. Therefore, even if (some) patent citations do not indicate spillovers as such, a reasonable interpretation is that they indicate a high probability of spillovers. Based on an American survey study, Jaffe *et al.* (1998) conclude that patent citations do indicate knowledge spillovers, but also that this indicator is a noisy one.

Patent citations have become a common measure of technology spillovers. Jaffe and Trajtenberg (2002) is a collection of important contributions. Several contributions have studied the geography of technology diffusion (Jaffe *et al.*, 1993) the time dimension of technology spillovers (Jaffe and Trajtenberg, 1996), the international dimension (Jaffe and Trajtenberg, 1999 and Sjöholm, 1996); the importance of universities in innovation (Jaffe and Trajtenberg, 1996); and measures of originality in knowledge production (Hall *et al.* 2002).

Trajtenberg (1993) investigates the *social* value of innovations in a specific technology field, CT-scanners, as a function of patent citations. Trajtenberg finds evidence that the social value increases with the number of subsequent citations to given patents. Several studies report results about correlations between the *private* value of patents and patent citations. Hall *et al.* (2005) study the market value of a sample of firms measured by patent counts weighted by patent citations. They find that citation weighted patents are positively correlated with firms' market values.

Maurseth (2005) use patent citation data to investigate technological rivalry and creative destruction. Maurseth finds that patent citations in general correlate positively with patent renewal (and therefore, probably, the value of patents). But patent citations within the same narrowly defined patent technology class (3-digit IPC) correlate with shorter survival of

patents. Maurseth interpret this as creative destruction of the cited patent from the citing patent.

In sum, innovation studies have incorporated patent citations as an important measure of technology flows. The best contributions have also discussed the weaknesses of this measure. The literature on patent citations in innovation research is surveyed in Jaffe and de Rassenfosse (2019).

Feng's paper

Sijie Feng rises further doubt about the usefulness of patent citations and proposes machine learning similarity indexes as her alternative. The main argument is the following: Since patent applicants add citations, they may be tempted to over-cite or under-cite. Therefore, there may be other reasons for the existence of citations than the technological relatedness between patents. If this is so, patent citations may be a noisy measure of technology spillovers (just as those who apply them admit they are). If the noise is not white but systematic, patent citations may be an even weaker measure.

Feng's strategy is to construct an alternative measure and compare her measure with patent citations. The alternative measure is machine leaning similarity indexes. Patent documents contain a detailed description of the invention in question. With the use of machine reading the essence of the patented technology can be extracted for each patent. Such data can in turn be used to evaluate similarity between all pairs of patents. Feng hypothesize that patents that are close by this measure are patents that are related to each other. As such Feng's technological distance measure is her alternative to patent citations.

Feng used her distance measure to evaluate the performance of patent citations. Her main findings are: i) non-local patents (i.e. applied for in different cities) are more likely uncited if they are very similar to each other relative to patents that are less similar, and ii) inventors tend to self-cite (cite their own inventions) less when they start working in new firms. These findings are Feng's main arguments for using similarity indexes from machine learning rather than patent citations as measure of technological relatedness.

Discussion

Feng uses machine learning to construct a measure of "textual similarity across pairs of patents as an alternative method of capturing knowledge relatedness across patents". She demonstrates that this measure of similarity is not parallel to patent citations measures. Feng argues that a possible reason for the discrepancy may be that patent citations "may be prone to strategic behaviour: inventors and firms may over-cite to offset the likelihood of litigation, and under-cite to increase the scope of the patent".

I agree with Feng that patent citations is probably a noisy measure of technological relatedness. This, however, is old news. Already Jaffe *et al.* (1998) concluded so. Feng proposes that her measure can serve as an alternative to patent citations. I only partly agree with her. It is probably correct that the usefulness of patent citations as a measure of technological relatedness is weakened if they are added for strategic reasons. As such, Feng's contribution is important. I think, however, that patent citations could fruitfully be complemented with Feng's measure.

However, the most interesting and important use of patent citations is not to indicate technological similarity. Rather they indicate some link of relevance between the cited and the citing patent. This link may be of different natures and can probably not be substituted with Feng's measure of proximity. The research on patent citations hypothesize and indicates that:

- Patent citations indicate some knowledge externality from the cited to the citing patent. The idea is that the cited patent may have stimulated the research and development that led to the citing patent. Similarity may not be important for such externalities. Patent citations indicate a link of relevance. It is clear that this link is noisy and sometimes non-existent. For Feng's measure, it is not clear that there is any link of relevance and therefore not necessarily and knowledge spillover.
- In line with the above, it may be that patent citations are indicative of the value of the cited patent. If the cited pated have external effects on subsequent research, citations may indicate social value. If the citations indicate that many want to imitate the cited patent, they may indicate private economic value. Trajtenberg (1993) (and many more) gave support to the first hypothesis. Hall *et al.* (2005) (and many others) gave support to the second. Feng's similarity index can hardly be indicative for private or social value. Maurseth (2005) proposes that patent citations may indicate technological rivalry against the cited patent. Feng's measure can hardly be used for such purposed.
- Jaffe and Trajtenberg (2002) demonstrate how patent citations can be indicative of generality. Their generality measure reflects the degree to which a patent is cited across technology classes. As such Feng's measure cannot be an alternative since her measure is similarity. Generality is the opposite.
- Jaffe and Trajtenberg (2002) proposes that the spread of citations made over technology classes indicates the citing patent's originality. Again, this measure is the opposite of similarity.

I agree with Feng that her measure indicates some weaknesses with patent citations. Her measure may potentially be a supplement to citation-based measures of similarity. Still, citations-based measures seem to one, out of few, measures of technological relevance between cited and citing patents. They will probably prove useful for this purpose also in the future.

References

- Hall, B., A.B. Jaffe and M. Trajtenberg (2002). The NBER Patent-Citations Data File: Lessons, Insights, and Methodological Tools. Ch. 13 in Jaffe and Trajtenberg (2002).
- Hall, B. A. B. Jaffe and M. Trajtenberg (2005). Market Value and Patent Citations. *Rand Journal of Economics* 36, 16-38.
- Haavelmo, T. (1954). *A Study in the Theory of Economic Evolution.* Amsterdam. North Holland Publishing Company.
- Jaffe, A. B., M. S. Fogarty and B. A. Banks (1998). Evidence from Patents and Patent Citations on the Impact of NASA and other Federal Labs on Commercial Innovation. *Journal of Industrial Economics,* 44, 183-205.
- Jaffe, A. B and de Rassenfosse, G. (2019). Patent citation data in social science research: overview and best

practices. In Depoorter, B., P. Menell and D. Schwartz (eds.) *Research Handbook on the Economics of Intellectual Property Law*, Vol. 1, Edward Elgar Publishing.

- Jaffe, A.B. and M. Trajtenberg (2002). *Patents, Citations and Innovations – a Window on the Knowledge Economy.* Cambridge Massachusetts, MIT Press.
- Jaffe, A. B. and M. Trajtenberg (1999). International Knowledge Flows: Evidence from Patent Citations.*Economics of Innovation and New Technology*, 8, 105-136.
- Jaffe, A. B. and Trajtenberg (1996). Flows of Knowledge from Universities and Federal Laboratories: Modelling the Flow of Patent Citations over Time and across Institutional and Geographic Boundaries. *Proceedings of the National Academy Sciences*, 93, 12671-12677.
- Jaffe, A. B., M. Trajtenberg and R. Henderson (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *Quarterly Journal of Economics,* 108, 577-598.
- Maurseth, P.B. (2005). Lovely but dangerous: The Impact of patent citations on patent renewal.*Economics of Innovation and New Technology* 14, no. 5: 351–74.
- Romer, P. M. (1986). Increasing Returns and Long-Run Growth.*Journal of Political Economy,* 94, 1002-1037.
- Sjöholm, F. (1996). International Transfer of Knowledge: The Role of International Trade and Geographic Proximity. *Weltwirthschaftliches Archiv,* 132, 97-115.
- Trajtenberg., M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovations.*RAND Journal of Economics*, 20, 172-187.