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

Le Ma[1]

1 University of Technology Sydney

Potential competing interests: No potential competing interests to declare.

This study uses unsupervised machine learning algorithm Doc2Vec to compare the text similarity of patents. The author proposes it is a superior measure of knowledge spillover than citations. Using the measure, the paper makes several empirical findings which suggest patent inventors' strategic patent citing behaviours. I appreciate the opportunity to review this paper and hope my comments will help improve the article.

First, the author uses a target patent's backward citations to identify other patents that are potentially citable by the target patent. Then the text similarity between the target and the potentially citable patents is compared pair-to-pair. My concern is whether these "potentially citable" patents are reasonably identified. The paper itself discusses many drawbacks of patent citations including underciting and overciting issues. This means that using common backward citations may wrongly match with irrelevant patents (if target patents overcite) or miss matching relevant patents (if target patents undercite). This makes the text-based similarity measure as unreliable as citations. A better approach would be to compare the text similarity for every pairs of existing patents which is expectedly a large task. But the paper should at least talk about the limitation and cautions readers about the findings. The author can also test the reliability of the measure by focusing on analysing all patents in a relatively small and specific tech field.

Second, the nature of patents vary in terms of knowledge generality and originality which leads to different knowledge spillover patterns. For example, a patent that develops basic scientific knowledge with high generality may be cited by patents from broader tech areas and more likely to have forward citations that share dissimilar text. The analysis doesn't control for such patent characteristics.

Third, the paper is descriptive in nature and the observed differences between local/non-local citations and before/after firm change are not statistically tested. It is unclear whether the differences are statistically or economically significant. The author can also use multivariate regression analysis to provide more insights.

Fourth, the paper compares only the patent abstract similarity not the full text which may not fully represent the content of a patent. This is in contrast to Acikalin et al., (2022). The author should provide examples to justify why it is reasonable to only look at abstract.

Fifth, the paper should provide examples of patent pairs with high and low similarity scores so that readers can get a sense about whether the similarity score differences are economically meaningful. Do patent texts significantly differ if similarity is 0.5 or 0.6?

Lastly, the paper should provide details of the empirical method. For example, it is unclear how the author empirically identifies inventors who change firms or inventors who move to other cities.

Reference:

Acikalin, U.U., Caskurlu, T., Hoberg, G. and Phillips, G.M., 2022. *Intellectual property protection lost and competition: An examination using machine learning* (No. w30671). National Bureau of Economic Research.