Research Proposal: A Machine Learning Analysis of Patent Screening with Implications for Productivity Losses from Inefficient Screening

Xiang Zheng*

April 2019

Abstract

In this proposal, I plan to analyze whether the current patent examination process in the US may be improved in terms of granting only better quality patents, leading to an improved allocation of economic resources. In particular, I want to develop a machine learning algorithm to re-evaluate each patent application under the current patent examination process. Based on my machine learning analysis, I can measure welfare (productivity) loss resulting from inefficient patent screenings by the operating performance, stock market performance and survival rate of firms owing these patents. In terms of research strategies, I plan to use unsupervised algorithms (such as natural language processing) to analyze unstructured data on application text and supervised algorithms (such as boosting trees and LASSO) to make predictions of application quality. I plan to use R or Python for this project.

*Ph.D. Candidate, Finance Department, Fulton 154B, Carroll School of Management, Boston College, Chestnut Hill, MA 02467, Tel: (617) 552-2049, Email: zhengxr@bc.edu.

For helpful comments and discussions, I thank my dissertation advisor, Thomas J. Chemmanur.

1 Motivation and Objectives

The patent system is viewed as one of the most important institutions that provides firms with innovation incentives by granting them temporary monopoly rights over their inventions. This, in turn, contributes to technological growth in the economy (see, e.g., Nordhaus (1969), Arrow (1972), and Mansfield (1986)). However, there has been considerable criticism of the patent system: critics argue that inefficient screening of patents reduce, instead of increasing, firms' incentives to innovate (see, e.g., Cornelli and Schankerman (1999), Lemley and Shapiro (2005), Schankerman and Schuett (2016), and Bessen and Maskin (2009)). One of the reasons for inefficiencies in the screening of patents may be the drastically increasing number of patent applications filed at the U.S. Patent and Trademark Office (USPTO). With a limited number of patent examiners available, the USPTO has been criticized for granting too many low-quality patents through a less scrutinized screening process (see, e.g., Heller and Eisenberg (1998), and Jaffe and Lerner (2011)). One potential economic consequence of inefficient patent screening is that it may mislead investors, thereby leading to misallocation in the economic resources invested in firms with low-quality patents (Cohen, Diether, and Malloy (2013)).

Motivated by the above criticisms of the existing patent screening system in the U.S., the existing theoretical and legal literature has explored the question of how to improve the patent screening system. In this vein, some have suggested changing the design of patent rights (see, e.g., Scotchmer (1999), and Hopenhayn and Mitchell (2001)); or changing patentability requirements (see, e.g., Dreyfuss (2008), and Kou, Rey, and Wang (2013)). In this proposal, I depart from the above literature and analyze how machine learning techniques may be able to improve the efficiency of the patent screening process through more efficient information processing (without changing the current patent system itself). As discussed above, each patent examiner faces increasing time constraints due to the growing number of patent applications filed each year, resulting in less rigorous scrutiny of patent applications (i.e.,

searching for prior art) resulting in the grant of more lower-quality patents (see, e.g., Lemley and Sampat (2012), and Frakes and Wasserman (2017)). On the other hand, the machine learning techniques I suggest here will utilize high computational power to analyze large-size (unstructured) data in a much faster way, thus increasing the efficiency of the patent screening system. In terms of prediction performance, machine learning generally makes much more accurate (out-of-sample) predictions by imposing fewer restrictions on the prediction function form compared to traditional econometrics, as I will show here in the context of my suggested machine learning algorithm.

2 Empirical Strategy and Computing Language

The empirical design to analyze the efficiency of patent screening process will follow three steps.

First, I will analyze unstructured data on application text and turn them into structured features using an unsupervised machine learning algorithm such as natural language process (NLP). The analysis in this step requires significantly large computation power so that the resources provided in this workshop will be very helpful.

Second, I will train a supervised algorithm (such as LASSO or Random Forests) using a training set to produce a prediction function that maps patent application characteristics (combining existing features such as backward citations and features generated from text data) to application quality.

Third, I will depart significantly from the standard supervised machine learning practice: in addition to merely measuring the accuracy of out-of-sample predictions using a test set, I focus on how my algorithm's prediction gives me information about the efficiency of actual examiners, and further measure the welfare (productivity) losses caused by inefficient patent screenings.

I plan to use R or Python for this project.

References

Arrow, Kenneth Joseph (1972), "Economic welfare and the allocation of resources for invention." In *Readings in Industrial Economics*, 219–236, Springer.

Bessen, James and Eric Maskin (2009), "Sequential innovation, patents, and imitation." *The RAND Journal of Economics*, 40, 611–635.

Cohen, Lauren, Karl Diether, and Christopher Malloy (2013), "Misvaluing innovation." *The Review of Financial Studies*, 26, 635–666.

Cornelli, Francesca and Mark Schankerman (1999), "Patent renewals and r&d incentives." *The RAND Journal of Economics*, 197–213.

Dreyfuss, Rochelle Cooper (2008), "Nonobviousness: a comment on three learned papers." *Lewis & Clark L. Rev.*, 12, 431.

Frakes, Michael D and Melissa F Wasserman (2017), "Is the time allocated to review patent applications inducing examiners to grant invalid patents? evidence from microlevel application data." *Review of Economics and Statistics*, 99, 550–563.

Heller, Michael A and Rebecca S Eisenberg (1998), "Can patents deter innovation? the anticommons in biomedical research." *Science*, 280, 698–701.

Hopenhayn, Hugo A and Matthew F Mitchell (2001), "Innovation variety and patent breadth." *RAND Journal of Economics*, 152–166.

Jaffe, Adam B and Josh Lerner (2011), *Innovation and its discontents: How our broken patent system is endangering innovation and progress, and what to do about it*. Princeton University Press.

Kou, Zonglai, Patrick Rey, and Tong Wang (2013), "Non-obviousness and screening." *The Journal of Industrial Economics*, 61, 700–732.

Lemley, Mark A and Bhaven Sampat (2012), "Examiner characteristics and patent office outcomes." *Review of Economics and Statistics*, 94, 817–827.

Lemley, Mark A and Carl Shapiro (2005), "Probabilistic patents." *Journal of Economic Perspectives*, 19, 75–98.

Mansfield, Edwin (1986), "Patents and innovation: an empirical study." *Management science*, 32, 173–181.

Nordhaus, William D (1969), "An economic theory of technological change." *The American Economic Review*, 59, 18–28.

Schankerman, Mark A and Florian Schuett (2016), "Screening for patent quality: Examination, fees, and the courts."

Scotchmer, Suzanne (1999), "On the optimality of the patent renewal system." *The RAND Journal of Economics*, 181–196.