Midterm Report

Overview

Innovation is singled out as the key factor that drives long-term productivity and economic growth. Countries that innovate, create new technologies and encourage the adoption of these new technologies grow faster than those that do not. That being said that, patents are the most commonly used measure of innovation output. Economists and historians mutually agree that data and statistics on patents act as an effective measure for technological change. In fact, the U.S. patent office (USPTO) issued its first patent on July 31, 1790. Since then, there have been over 6 million patents issued leading to breakthrough development of steamships, automobiles, electric power, electric appliances, aviation, aerospace, telecommunications, mobile communications, computers, the Internet, biotechnology, nanotechnology, Machine Learning, Natural Language Processing, Artificial Intelligence, and blockchain. I aim to explore the patent ecosystem using the data provided by USPTO and PatentsView to understand what factors drive patent innovation (thereby understand the drivers for technological and economic growth) as part of this project. USPTO and PatentsView have undertaken commendable efforts to document patent data (both recent patents and those filed as early as the 1970s). Such extensive and exhaustive datasets help us understand the geography of innovation, demography of inventors and factors (if any) that inhibit patent innovation. It is important to learn these given the tremendous personal and commercial benefits that can be reaped by patenting. Using the above mentioned data sources, I uncover essential trends in innovation over a broad time span and across different categories, such as drugs and medical, computer and communications, etc.

For each patent application, I identify diverse characteristics, such as the number of inventors, gender of inventors and various issuance and post-issuance outcomes (details below). The datasets used also identify the primary and secondary technology class in the United States Patent Classification (USPC) system to which the application was assigned.

About the Data

- **USPTO research datasets**: https://www.uspto.gov/ip-policy/economic-research/research-datasets: The United States Patent and Trademark Office (USPTO) is an agency in the U.S. Department of Commerce that issues patents to inventors and businesses for their inventions, and trademark registration for product and intellectual property identification.
- **PatentsView**: https://patentsview.org: PatentsView is a free, online platform for visualizing, disseminating, and promoting a better understanding of U.S. patent data supported by the USPTO's Office of the Chief Economist

Both the datasets were straightforward to merge using the patent application number, which is a unique universal identifier for an application, with minor pre-processing steps, such as :

- Removing the leading '0' (appended to primarily have a uniform length for the application numbers) and converting string application numbers to integers
- Aggregating the patent and inventor file at the patent-level and computing the proportion of female and male for these patents

- Evaluating the number of days to review a patent application by computing the difference in days between the grant date and filing date of a patent application (computed for ~1.6 million accepted patent applications only)
- Aggregating the total forward citations that a patent application has received from the patent ID1, patent ID2 total citations file (computed for ~1.6 million accepted or granted patent applications only)
- On further analysis, I realise that the data is most consistent across 2001-2014 (in terms of missing values and availability of features such as citations). Hence, the results are restricted from 2001 to 2014, for now.
- I also restrict the applications to those for which I have been able to successfully merge the gender dataset from PatentsView.

Variables used in the analyses

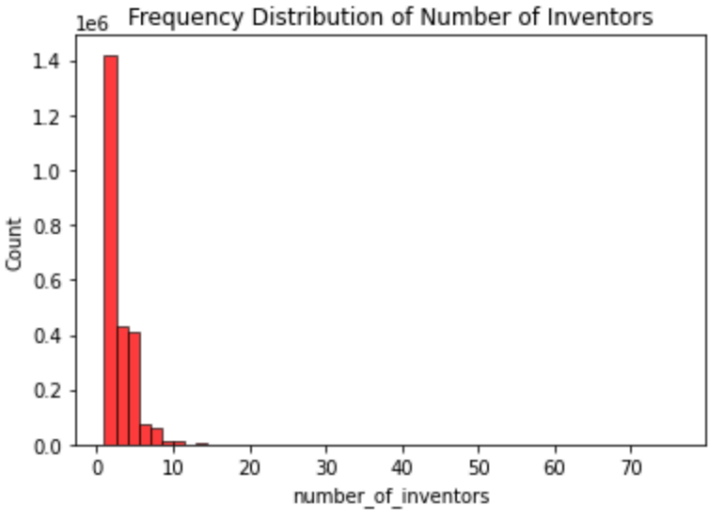
- **Patent issued**. An indicator variable that equals a value of one if the patent requested in the application is granted.
- **Days between filing and issuance**. The length of the review process, estimated as the number of days between the date of filing of the patent application and the date of issuance of the requested patent.
- **Total citations**. Total citations received by a granted patent from patents granted in the future constitute an effective proxy for a patent's value and a strong indicator of innovation (Hall et al., 2005). Citations can be added by examiners, applicants, lawyers, etc.
- **Proportion of female inventors**.
- **USPC primary class.** USPTO assigns new patent applications to one of more than 400 general (primary) technology class and to one or more subclasses. Each class and subclass is identified by a unique code (Graham et al., 2015). Because the patent review process can vary significantly for different technological classes, all models control for an application's primary technology class.
- **Application filing year.** All models include fixed effects for the filing year of an application.

It is important to incorporate classification and year information will control for any confounding effects on patent outcome due to its classification and year of filing. It prevents the model from under-fitting and provides for a complex enough hypothesis for estimation. Moreover, this report is concerned with estimation and not prediction (that is, for a new patent, I do not intend to predict whether it will be granted), hence, we need not be concerned about overfitting. I intend to extend my analysis with inventor-, examiner- and organisations (filing these patents) -level features such as inventor experience, the type of assignee/organisation (small, large, etc.).

Descriptive Summary

As PatentsView provides the gender of each inventor using an involved and accurate disambiguation technique, I deep dive into the distribution of the inventors across patent applications and also report a descriptive summary for all variables across 2.4 million observations (patent applications) in Table 1 below.

Figure 1 and 2: Below show the distribution of total number of inventors across all patent applications and years. We also show the distribution by gender.



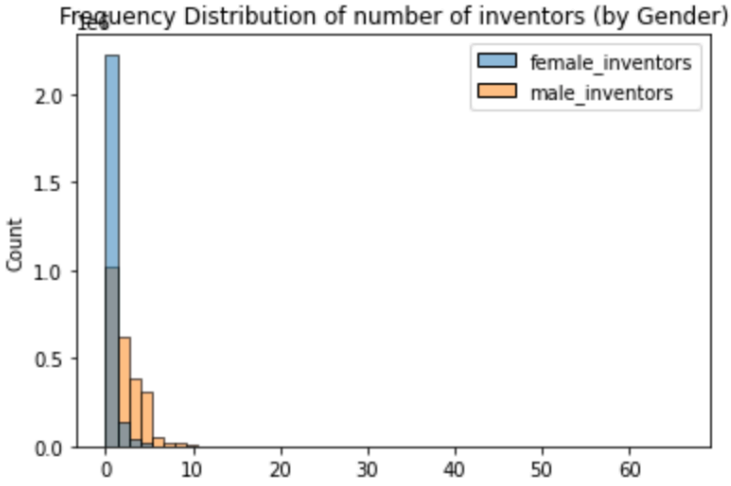


Figure 3: Below zooms into some trends showing average proportion of female inventors across patent applications by year and by varied technological categories provided by NBER, such as Communications, Drug, etc. We see a steady increase in participation by women inventors across varied categories.

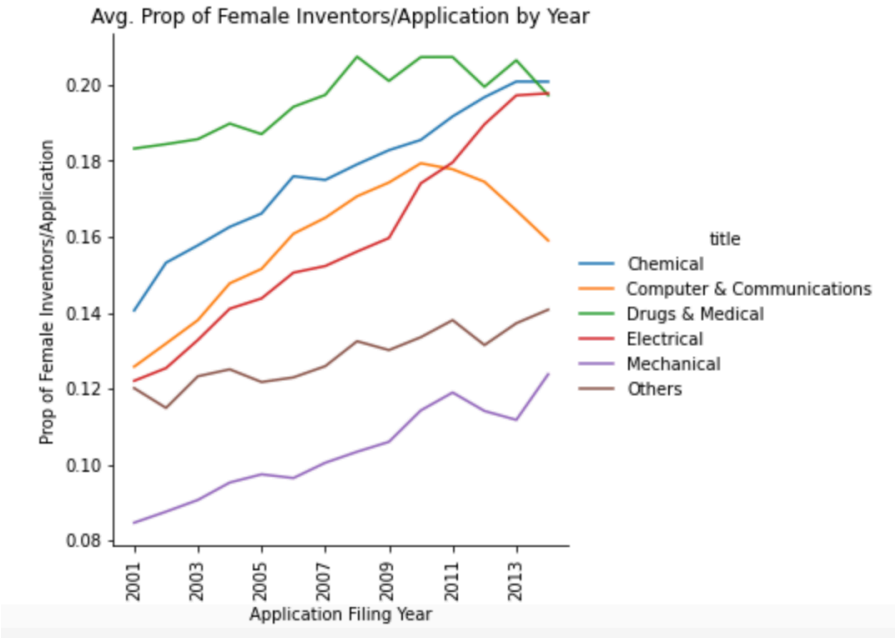


Figure 4: Below provides a wider perspective in invention rates by gender and year. This figure illustrates a stark gap in the average proportion of female and male inventors across years, despite the increase invention rates by women inventors shown in Figure 3 above.

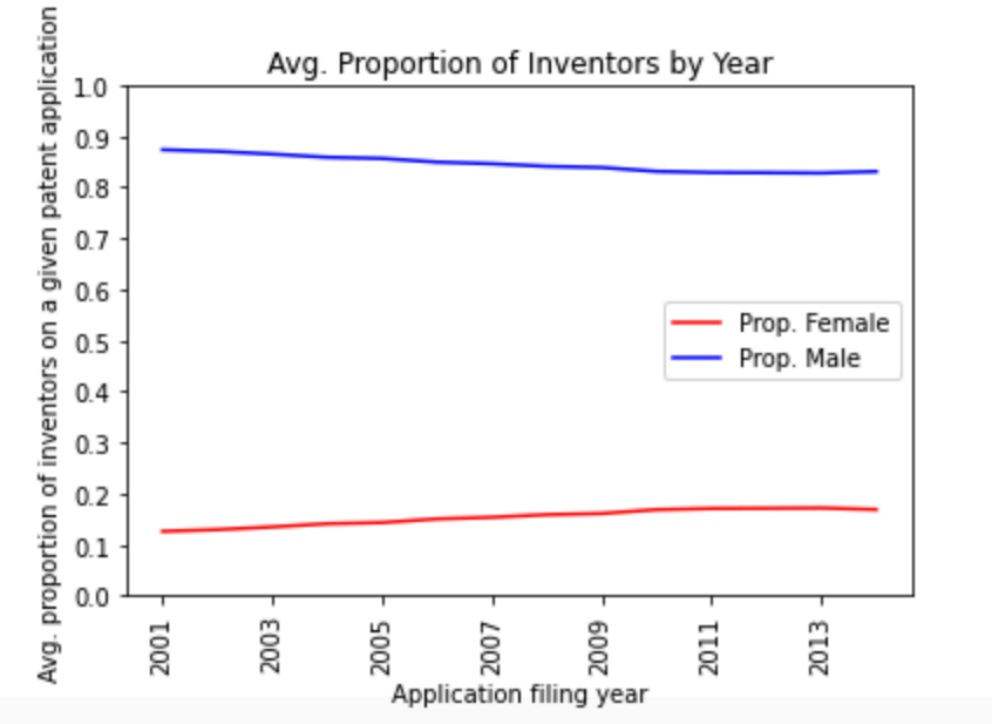


Table1: Descriptive summary for different variables collated from USPTO and PatentsView. The table 1 above provides a summary for 2,425,468 patent applications

	Mean	**Standard Deviation**	**Min**	**Max**
Patent issued	0.679	0.467	0.000	1.000
Days between filing and issuance	808.331	733.770	0.000	5937.000
Total citations	11.429	35.827	1.000	3439.000
Prop. female	0.436	0.416	0.000	1.000
Female Inventor Count	1.396	1.689	0.000	46.000
Male Inventor Count	1.772	1.795	0.000	44.000
Filing Year	2007.418	3.274	2002.000	2014.000

Estimation Approach

I now concentrate my study to analyse the gap in invention observed in Figure 4 by estimating the effects of Proportion of female on varied patent outcomes described above. As part of this analysis, I attempt to model the different outcomes, using linear regression or logistic regression, based on which fits my data better. Both these models can be written as:

$$Y_j = f(Prop.Female_j, \tau_j, \gamma_j)$$

where Y_j is the outcome of interest, Prop. Female is the proportion of female inventors on the applicant team for the application j, τ_j is a vector of technology class fixed effects, and γ_j is a vector of filing year fixed effects.

- **Patent issued**. Estimated using a logistic regression classifier, since patent issued is encoded as a categorical variable.
- **Days between filing and issuance**. Modelled using linear regression with least squared estimates.
- **Total citations**. Will be modelled using linear regression with least squared estimates.

The final model and results have been reported in Table 2 below.

Table2: Provides the estimated effects of Prop. Female on Patent Issued and Days to grant. I also report the goodness of fit of the model (R^2) and standard errors, p-value of my point estimates.

	Patent Issued	**Days between filing and issuance**
Prop. female	-0.0615***	0.0231***
	(0.00363)	(0.000839)
Constant	0.104	7.146***
	(0.359)	(0.142)
Observations	24,25,468	16,47,162
R-squared	0.0721	0.240
Log Likelihood	-1.412E+06	-1.710E+08

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Conclusion and next steps

To dive deeper into the statistical significance of the gap illustrated in Figure 4, I analyse two outcome variables - patent acceptance rate and the number of days taken to review a filed application - in Table 2. We see that Prop. Female is negatively associated with patent issue probability, which implies that the probability of a patent being issued reduces as the proportion of female inventors increases. We also see that the number of days to grant a patent application increases as the female inventors increase (since it is positively associated with the outcome). These results align with those in several notable literature Jensen et al., 2018, Miguelez et al., 2019, among others. Our results indicate (with very high confidence) that the current patent examination system imposes a penalty on patent applications with higher proportion of female inventors. In a sense, this may hinder technological growth due to reduced participation of women inventors in the innovation ecosystem.

Over the semester, as the project progresses, I aim to extend the analyses to also include Total citations as an outcome variable in the analysis. I also intend to add other covariates, such as the

type of organisation filing the patent, examiner experience, inventor experience, etc., that may help explain some variance in pre- and post-examination outcomes. These may also help in controlling for any confounding effects due to these and under-fitting. Furthermore, I intend to test these results by performing estimation on various subsamples of the data and using different estimation techniques like ridge and lasso.

References

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Graham, Stuart J.H. and Marco, Alan C. and Miller, Richard, The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination (November 30, 2015)

Jensen, K., Kovács, B., & Sorenson, O. (2018). Gender differences in obtaining and maintaining patent rights. *Nature biotechnology*, *36*(4), 307-309.

Miguelez, E., Toole, A., Myers, A., Breschi, S., Ferruci, E., Lissoni, F., ... & Tarasconi, G. (2019). *Progress and Potential: A profile of women inventors on US patents* (No. hal-02274254).