Final Project Report

Motivation & Background

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. Over the course of analysis, I also narrowed the project to assessing the demography of inventors in the patent ecosystem.

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

Data Preparation

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 total citations). Hence, the results are restricted from 2001 to 2014
- I also restrict the applications 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.
- **No. of inventors.** The size of the inventor team can significantly impact the quality of a patent application and hence affect its outcome (Jones et al., 2007). I therefore control for the number of inventors on an application.
- **Small entity.** An indicator variable set to 1 if the applicant qualifies as a small entity. A small entity is typically either an individual inventor, a collaboration of individual inventors, a non-profit organisation, or a company with fewer than 500 employees. Small entity status typically entitles the applicant to a 50 percent discount on most fee payments to PTO (Graham et al., 2015).
- **Inventor experience.** Logged average of the number of applications filed previously by all inventors who authored a focal application.
- **Examiner experience.** Logged count of patent applications handled by the examiner prior to the focal application. Examiners' prior experience can significantly impact the review process and outcomes of an application.

Addressing underfitting and overfitting

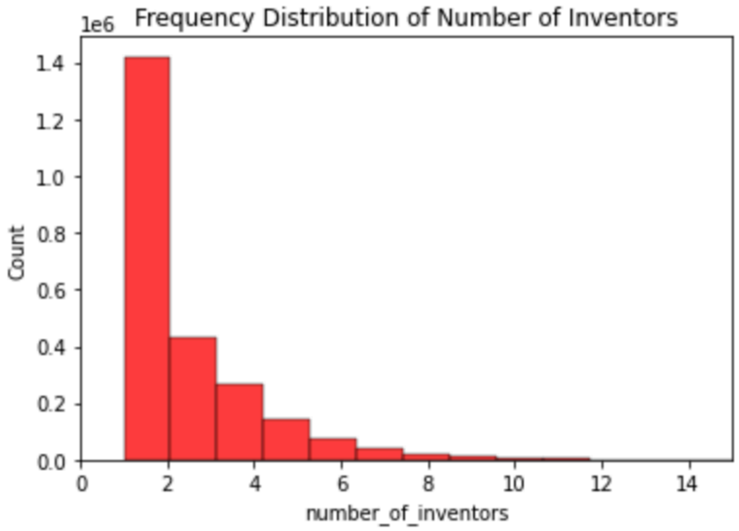
It is crucial to incorporate patent class and year information to control for any confounding effects of patent category and filing year on these patent outcomes. For instance, in a given year, say 2000, we observe an increase in the patents filed (probably due to the WWW and internet boom), we will naturally observe higher patent acceptance, rejection rates and a higher count of days to review these patent applications. Thus, we add all distinct patent classes and filing years as covariates in our model (as nominal variables). Furthermore, it prevents the model from under-fitting and provides for a complex hypothesis for estimation.

Moreover, the dataset in this project comprises a representative sample of all patents filed from 2001 to 2014. Hence, this report is intended to provide an accurate estimation of the coefficient of female coefficient and not prediction (that is, for a given patent, I do not intend to predict whether it will be granted since we already know its outcome). Hence, we are primarily concerned about getting accurate point estimates for each of the covariates described above and not about overfitting. In some sense, we want our models to overfit to the data and give us an estimate as close to the true estimate (since our data is the entire population of patent applications). Thus, we try to model our estimates such that they provide the best fit to the dataset without worrying about overfitting.

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.



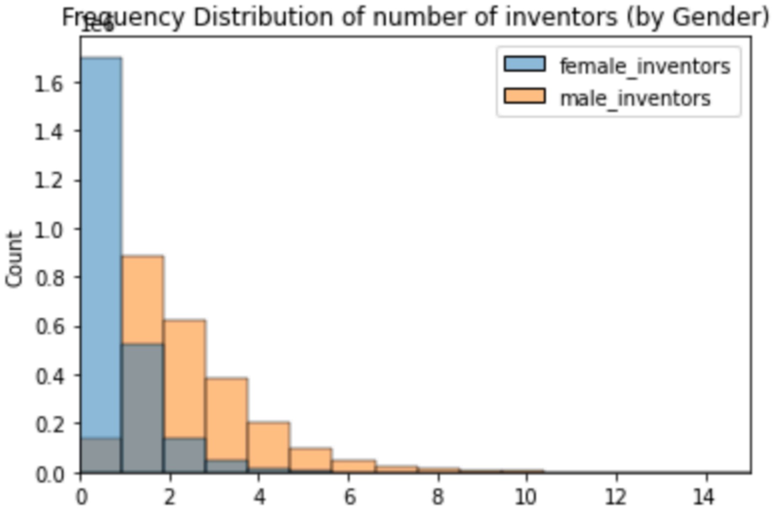


Figure 3: Below zooms into some trends showing the mean proportion of female inventors across patent applications by year and by varied technological categories provided by NBER, such as Communications, Drugs, etc. We see a steady increase in participation by women inventors across varied categories.

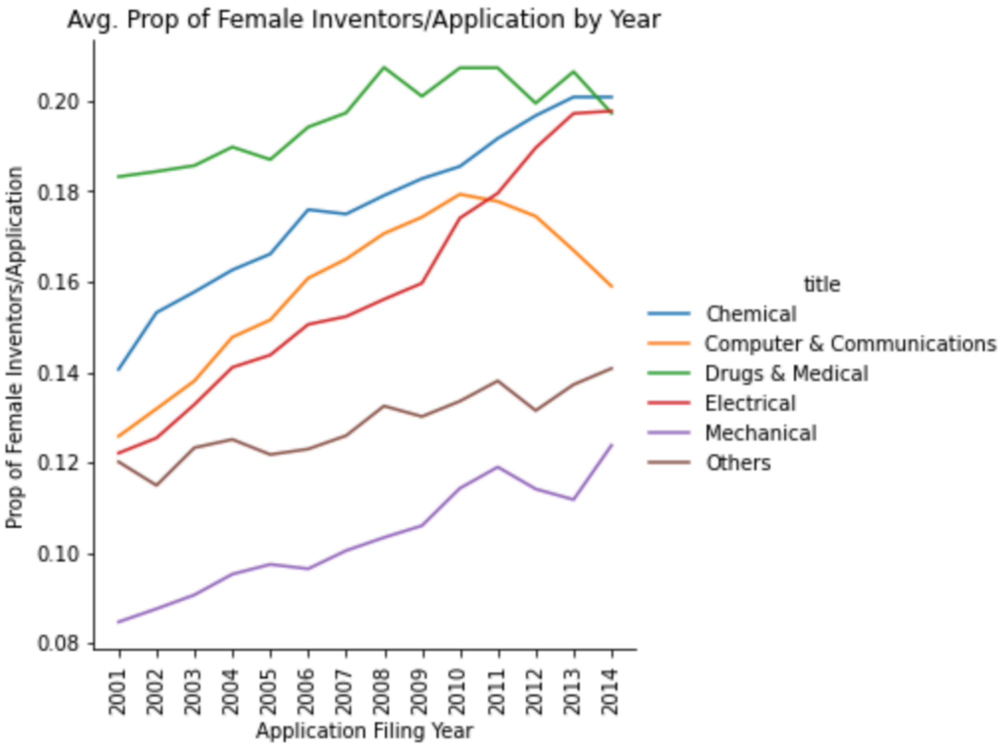


Figure 4: Below provides a wider perspective in invention rates by gender and year. This figure illustrates a stark gap in the average proportion of female and male inventors across years, despite the increase invention rates by women inventors shown in Figure 3 above.

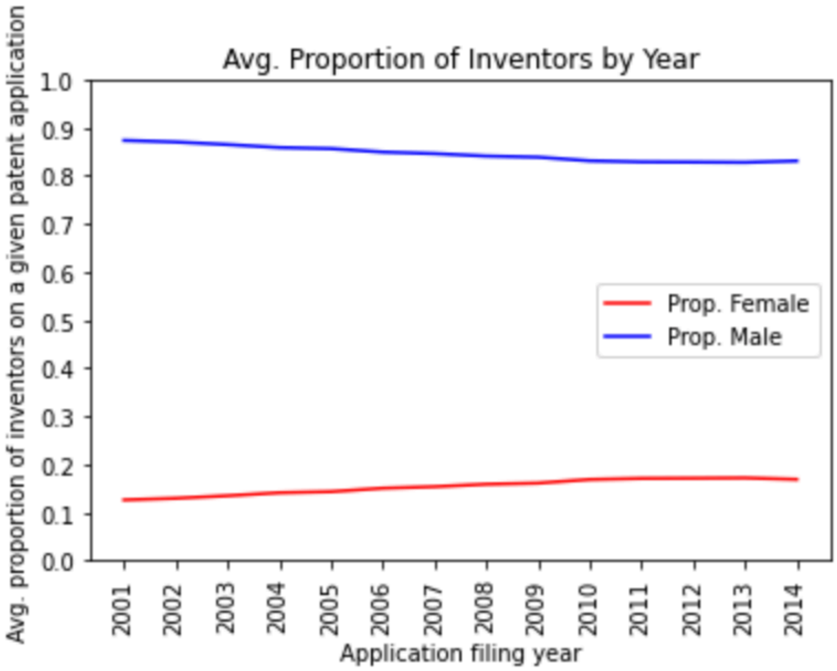


Table1: Descriptive summary for different variables collated from USPTO and PatentsView. The table 1 below provides a summary for 2,425,468 patent applications

	Mean	**Standard Deviation**	**Min**	**Max**
Patent issued	0.679	0.467	0.000	1.000
Days between filing and issuance	808.331	733.770	376.00	5937.000
Total citations	11.429	35.827	1.000	3439.000
Prop. female	0.236	0.416	0.000	1.000
Inventor experience	13.376	61.388	1.000	3280.000
Examiner experience	242.883	221.057	0.000	1700.000
Small entity	0.232	0.422	0.000	1.000
Filing Year	2007.418	3.274	2001.000	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**. Modelled using linear regression with least squared estimates.

What is being tested?

Null Hypothesis:

- Coefficient for proportion of female inventors = 0

Alternative Hypothesis:

- Coefficient for proportion of female inventors < 0 for patent grant rate and citations
- Coefficient for proportion of female inventors > 0 for days to review

Table 2: Provides the estimated effects of Prop. Female on Patent Issued, Days to grant and Total citations. 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**	**Total (forward) citations**
	Logistic Regression	Linear Regression	Linear Regression
Prop. female	-0.0615***	0.0231***	-0.654***
	(0.00363)	(0.000839)	(0.0139)
Constant	0.104	7.146***	1.313***
	(0.359)	(0.142)	(0.268)
No. Of Observations	24,25,468	16,47,162	11,77,642
Class, Year Fixed Effects	Yes	Yes	Yes
R-squared	0.0721	0.240	0.0422
Log Likelihood	-1.412E+06	-1.710E+08	-3.863E+06

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3: Provides the estimated effects of Prop. Female on Patent Issued, Days to grant and Total citations, with additional co-variates. 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**	**Total (forward) citations**
	Logistic Regression	Linear Regression	Linear Regression
Prop. female	-0.0583***	0.0211***	-0.127***
	(0.00414)	(0.000914)	(0.0134)
Team Size (Number of inventors)	0.0346***	0.00485***	0.0535***
	(0.000786)	(0.000151)	(0.00128)
Small Entity Indicator	-0.808***	-0.0305***	-0.00718
	(0.00386)	(0.000975)	(0.00530)
Examiner Experience	0.332***	-0.0927***	0.0493***
	(0.00157)	(0.000330)	(0.00219)
Inventor Experience (Logged)	0.109***	-0.00867***	0.0959***
	(0.00169)	(0.000340)	(0.00241)
Constant	-1.005**	7.453***	0.965***
	(0.411)	(0.125)	(0.214)
No. Of Observations	24,25,468	16,47,162	11,77,642
Class, Year Fixed Effects	Yes	Yes	Yes
R-squared	0.126	0.313	0.0509
Log Likelihood	-1.330E+06	-1.540E+08	-3.828E+06

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Analysing Table 2

In table 2, we observe that the estimate for female proportion is statistically significant. The p_value of coefficient of female proportion for each our outcome is less than 0.01, providing a strong evidence against null hypothesis. Thus, we can say that keeping everything else constant, for 1 unit increase in the proportion of female inventors on a given patent application, the log odds of a patent being accepted reduces by 0.0615 units, the number of days to grant increases by 0.0231 units and citations received by the patent reduces by 0.654 units.

Analysing Table 3

Similarly, in table 3 we see our estimate for female proportion is statistically significant, even after controlling for various patent, assignee, inventor and examiner level characteristics. The p-value of the coefficient for each our outcome is less than 0.01, providing a strong evidence against null hypothesis. Thus, we can say that (keeping everything else constant), for 1 unit increase in the proportion of female inventors on a given patent application, the log odds of a patent being accepted reduces by 0.0583 units, the number of days to grant increases by 0.0211 units and citations received by the patent reduces by 0.127 units.

Techniques applied

- Missing values imputation using mean for examiner and inventor experience
- Logistic and linear regression models
- Trends analysis

Conclusion and future work

In summary, the variables modelled indicate biases from various actors such as examiners (patent acceptance and days to review) and citations (other applicants) involved in the patent ecosystem. Our results indicate (with very high confidence) that the current patent examination system imposes a penalty on patent applications with a higher proportion of female inventors. These results align with those in several notable literature Jensen et al., 2018; Miguelez et al., 2019, among others.

Consequences of our finding: This penalty may hinder technological growth due to the reduced participation of women inventors in the innovation ecosystem. Hence, we should take concrete steps to anonymise the patent examination and citation process.

Future extension: We can try to incorporate the textual content of patents in our analysis and add better proxies for patent quality.

References

1. Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.
2. 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)
3. Jensen, K., Kovács, B., & Sorenson, O. (2018). Gender differences in obtaining and maintaining patent rights. *Nature biotechnology*, *36*(4), 307-309.
4. 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).
5. Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science,* 316(5827), 1036-1039.
6. Graham, S. J., Marco, A. C., & Miller, R. (2015). The USPTO patent examination research dataset: A window on the process of patent examination. Georgia Tech Scheller College of Business Research Paper No. WP, 43.