Pset1 : Data Section

MACS 30200 - Perspectives on Computational Research Luxi Han 10449918

Data Source

The data set that we use for this paper is from The NBER Patent Citations Data Files. This data set is curated by Hall, Jaffe and Trajtenberg (2011). The data set combines two data sources: patent data from Patent Office and firm level dataset from Compustat. The data can be accessed from the NBER website. [1]

The first data source is patents granted during January 1, 1963 and December 30, 1999 from United States Patent and Trademark Office (USPTO). This original data set consists of 10 variables and contains information on patent grant and application date, country of the inventor, assignee information, patent class, etc. The curator of the data set also constructs other new variables on techonological category, number of citations made and received, and two measures of generality and originality. Specifically, citations made refers to the citation one specific patent made and citations received refers to all of the citations one patent receive after being granted. Measure of generality represents how wide of the field one specific patent influence. Measure of originality is computed using citations made by one patent. Detailed construction methods are specified in the article *The NBER Patent Citations Data File: Lessons, Insights And Methodological Tools* [2]

The dataset has been used for a variety of paper and research fields. Hall, et al.(2005)[3] used this dataset to study the relationship between a firm's patents and the market value of the firm. They confirm that the more important of patents granted to a firm, the more market value a firm has. This indicates that innovation can boost the market value of a firm. The same dataset is also used to study patent related issues for countries other than the US. Hu and Jaffe(2001)[4] study knowledge diffusion among Asian countries by studying the inventors patented in the US.

Summary Statistics

In table 1, I show the summary statistics.

The dataset contains about 3 million observations. The patents are included for patents granted between 1963 and 1999. Number of citations made refers to the number of citations made in one specific patent. The average citations made is around 7,720 per patent. This number is used by Hall, et al. to consturct the originality measure of a patent. Number of patent received, which refers to number of times one patent is cited by other patents, has an average of 4.779 per patent. This measure in general represents the importance of a patent. In general, the more one patent is cited, the more influential and general the patent is.

Mean forward citation lag is defined as the average difference in years between one patent and patents citing it. While mean backward citation lag is defined as the average difference in years between one patent and patents it cited. The former has an average around 8.306 years and the latter has an average of 14.1 years. Notice that the forward lag is less than the backward lag. This is majorly because the patent cited are usually patents that are more influential and fundamental. But the forward lag is computed for citations for each patent. But most of the patents are not influential. This significantly shortens the forward lag since most of the patents is outdated in a short period of time.

[1] http://nber.org/patents/

[2] Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. The NBER patent citation data file: Lessons, insights and methodological tools. No. w8498. National Bureau of Economic Research, 2001.

[3] Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. "Market value and patent citations." RAND Journal of economics (2005): 16-38.

[4] Hu, Albert GZ, and Adam B. Jaffe. "Patent citations and international knowledge flow: the cases of Korea and Taiwan." International journal of industrial organization 21, no. 6 (2003): 849-880.

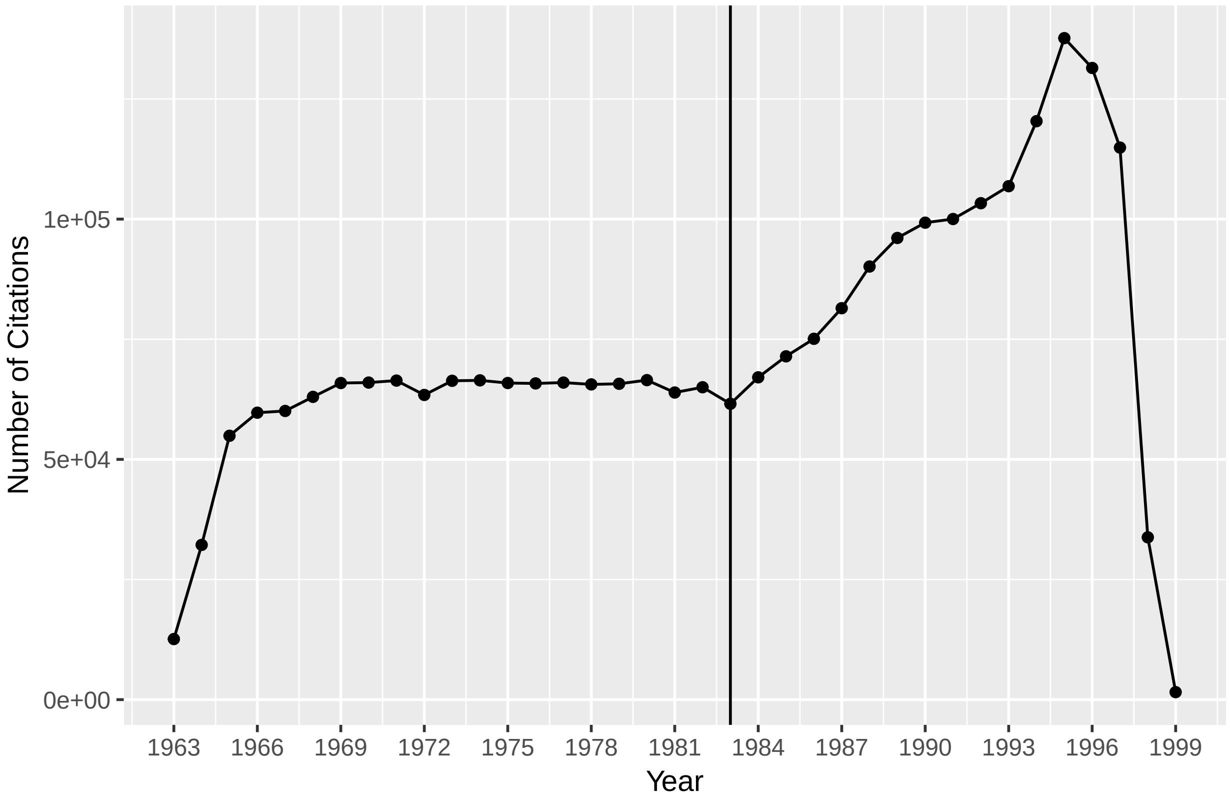
Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Grant Year	2,923,922	1,983.548	10.978	1,963	1,999
Grant Date	2,923,922	8,785.647	4,013.281	1,096	14,606
Application Year	2,699,606	1,983.106	10.128	1,901	1,999
Number of Claims	1,984,055	12.083	10.268	1	868
Number of Citations Made	2,139,314	7.720	9.000	0	770
Number of Citations Received	2,923,922	4.779	7.346	0	779
Measure of Generality	2,240,348	0.321	0.285	0.000	0.940
Measure of Originality	2,042,151	0.349	0.281	0.000	0.951
Mean Forward Citation Lag	2,074,641	8.306	5.804	0.000	96.000
Mean Backward Citation Lag	2,088,785	14.100	11.769	0.000	154.000
Share of Self-Citations Made-Upper Bound	1,703,004	0.136	0.256	0.000	1.000
Share of Self-Citations Made-Lower Bound	1,703,004	0.110	0.218	0.000	1.000
Share of Self-Citations Received-Upper Bound	1,599,160	0.132	0.260	0.000	1.000
Share of Self-Citations Received-Lower Bound	1,599,160	0.125	0.250	0.000	1.000

The curators construct generality and originality measure from the combined dataset using citations made and received. The construct method is specified in the following section.

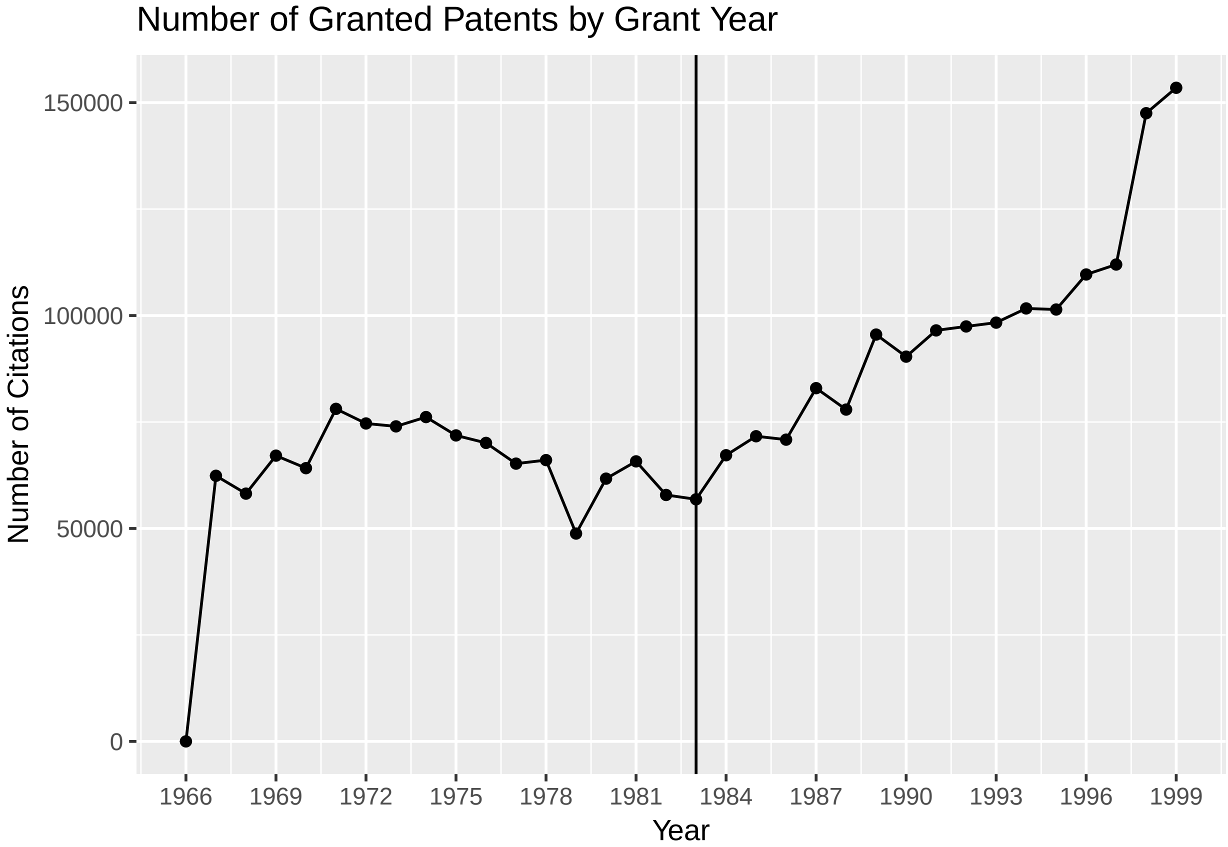
Now we show the trend of number of granted patents by application year. We can see that there is a clear trend before and after 1983. Before 1983, number of granted patents is rougly the same each year. After 1983, we see that there is a clear upward trend indicating there is an increasing number of patents granted each year. This is likely caused by the change in the granting process of the Patent Office in the 80s.

Number of Granted Patents by Application Year

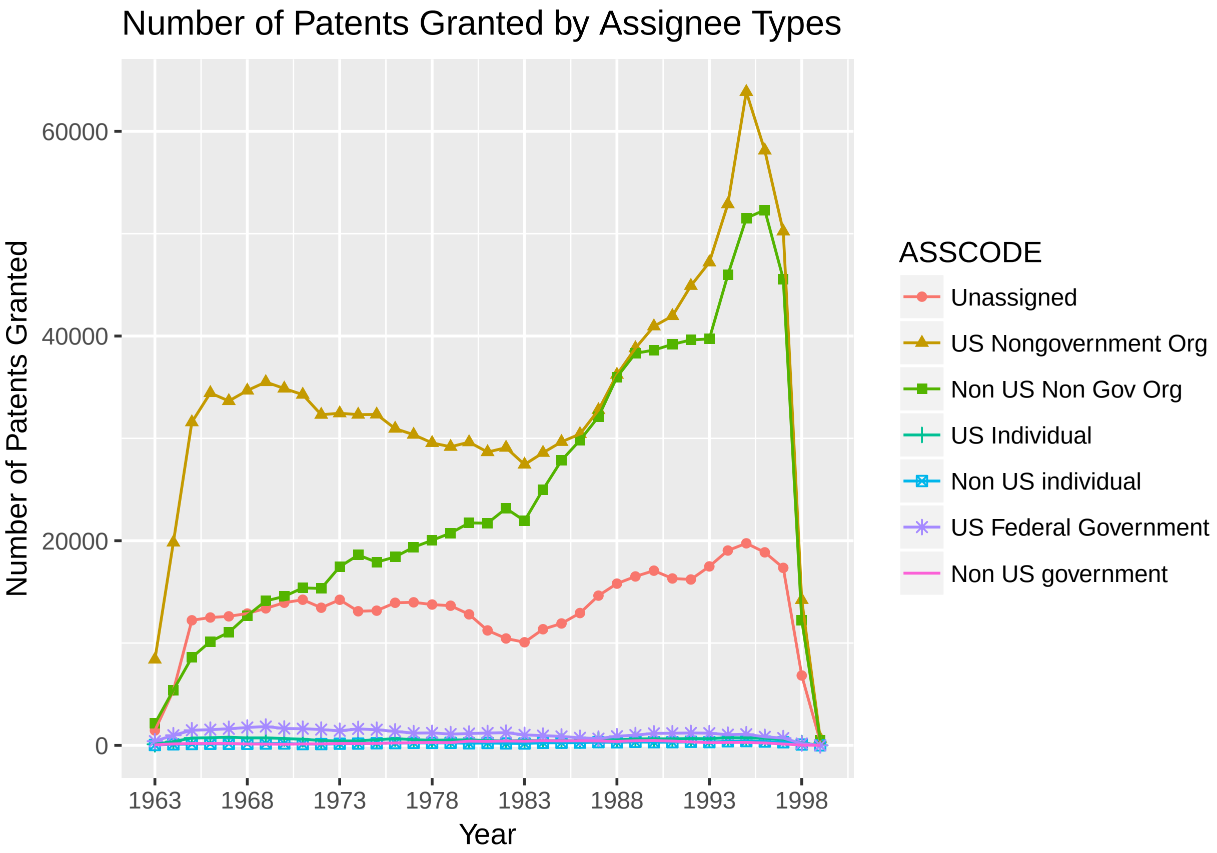


Furthermore, we notice that there is a drastic drop after 1995. This indicates the truncation problem. Normally, it will take around two years for the patent to be granted. Normally, around 95 percent of patents successfully granted will be granted in 3 years. Since we know the data period is from 1963 to 1999, then the drop after 1995 should be caused by the truncation problem. For example, patents filing application in 1998 will get the granting permission in 2000 but the data will not cover this part of patents.

To corroborate our argument, we plot the number of patents granted in granting year. The truncation problem disappear.



Now we can take a look at number of patents granted by assignee types. We can see that organizations file for most of the patents applications. Specifically, US non government organizations are granted most of the patents. This corresponds to reality, since corporations with R&D fundings fall in this category. Governmental agencies and individuals are granted for a bout the same number of patents.



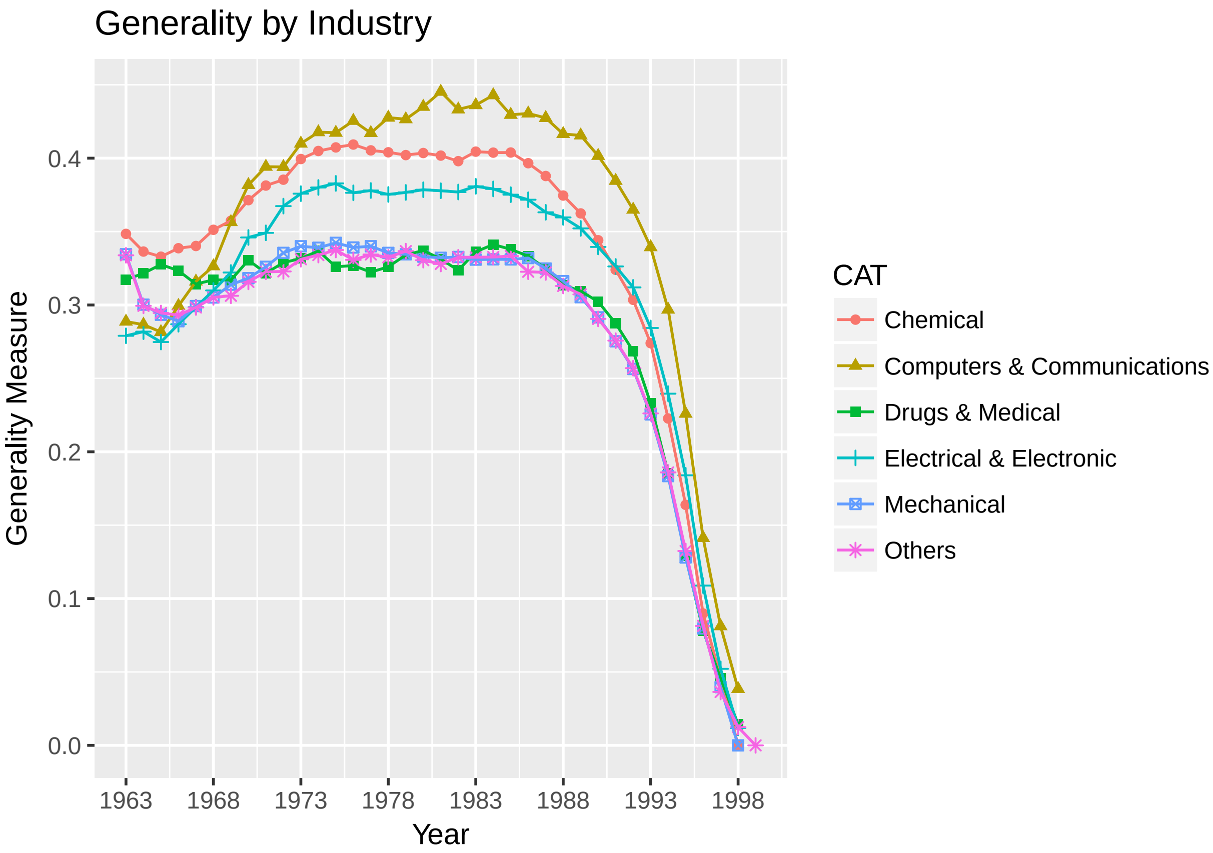
We then turn our attention to industry differences. Specifically, we delve into the generality and originality of different fields.

Firstly, generality is computed in the following fashion:

$$Generality_i = 1 - \Sigma_j^{n_i} s_{ij}^2$$

$s_i j$ refers to the percentage of citations received by patent i that belong to patent class j. Then the more fields cite this patent, the higher the generality measure is.

Then, we can see computer and communication industry has the highest generality. This is because computer technology is a common use techonlogy across different fields. Mechnical and medical industries are the ones with the least generality since they are mostly specialized fields.

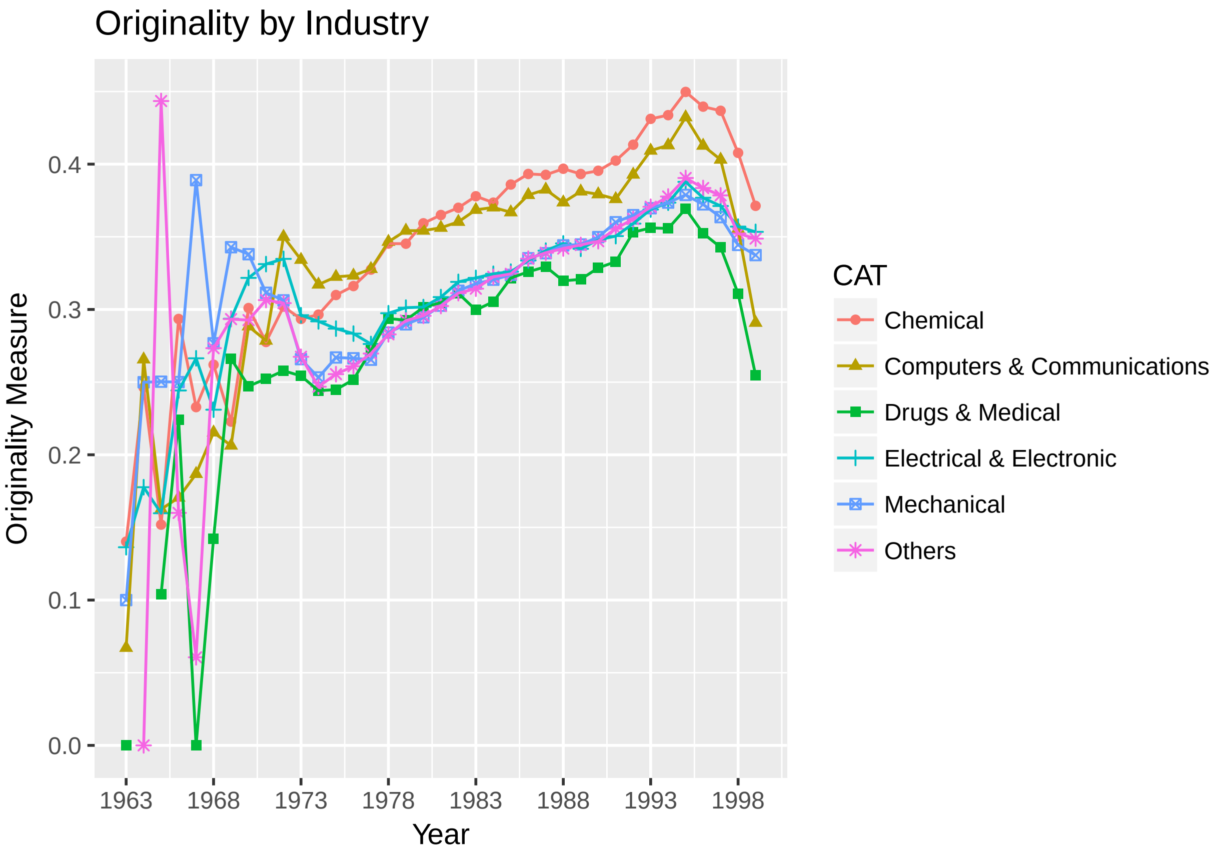


Then we can turn our attention to originality.

$$Originality_i = 1 - \Sigma_j^{n_i} h_{ij}^2$$

$h_i j$ refers to the percentage of citations made by patent i that belong to patent class j.

This is different from generality in that it uses citations made instead of received. Then the wider the patent involves, the higher its originality. This gives us mostly the same picture as originality. The originality is the highest for the chemical industry with computer and communication industry following at the second place.



Notice that there is also a suddent drop after 1995. Since the originality and generality measure is positively correlated with number of citations received. Then the sudden drop is also a reflection of the truncation problem.