Persp-Research PS1

Ningyin Xu

Due 4/19/2017

U.S. Patent Data

General Description

This NBER U.S. Patent Citations Data File is obtained from the National Bureau of Economic Research website, **http://nber.org/patents/**. The major database is listed in the third row in the table in the patent page, with the download links for documentation file **"pat63_99.txt"**, .tpt-formatted data **"pat63_99.zip"**, and csv-formatted data **"apat63_99.zip"**.
This data includes all utility patents in the USPTO's TAF database granted during the period 1963 to December 1999. The source of this data includes: United States Patent and Trademark Office (USPTO), and Hall, B.H., A. B. Jaffe, and M. Trajtenberg's computations (2001).

A detailed description of this data can be found in

> *Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.*

Further documentation on uses of the patent citation data, including the methodology paper and a CD containing the complete dataset itself, is available in the book *Patents, Citations and Innovations: A Window on the Knowledge Economy by Adam Jaffe and Manuel Trajtenberg, MIT Press, Cambridge (2002).* The book may be ordered from MIT Press. ISBN 0-262-10095-9.

Key papers (papers that have been cited the most times) that have used this data include:

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

> *Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.*

> *Acs, Zoltan J., Luc Anselin, and Attila Varga. "Patents and innovation counts as measures of regional production of new knowledge." Research policy 31.7 (2002): 1069-1085.*

The database is developed based on the patent records from USPTO, which has 10 original variables. Hall, B.H., A. B. Jaffe, and M. Trajtenberg computed and constructed 10 new variables, including technological category/sub-category, number of citations made/received, percent of citations made by this patent to patents granted since 1963, measure of "generality" and "originality", mean forward citation lag, mean backwards citations lag, and percentage of self-citations made –upper and lower bounds.

Descriptive statistics

Key Variables	Mean	Std	Median	Min	Max
Number of Citations Made	7.720	9.000	6.000	0	770.000
Number of Citations Received	4.779	7.346	3.000	0	779.000
Percent of Citations Made to Patents Granted Since 1963	0.843	0.249	1.000	0	1.000
Measure of Generality	0.321	0.285	0.375	0	0.940
Measure of Originality	0.349	0.281	0.420	0	0.951
Mean Forward Citation Lag	8.306	5.804	7.000	0	96.000
Mean Backward Citation Lag	14.100	11.769	10.500	0	154.000
Share of Self-Citations Made - Upper Bound	0.136	0.256	0.000	0	1.000
Share of Self-Citations Made - Lower Bound	0.110	0.218	0.000	0	1.000
Share of Self-Citations Received - Upper Bound	0.132	0.260	0.000	0	1.000
Share of Self-Citations Received - Lower Bound	0.125	0.250	0.000	0	1.000

Descriptive statistics table above can tell us some basic information about this dataset. Citations are informative of links between patented innovations. Citations made constitute a "paper trail" for spillovers, i.e. the fact that patent B cites patent A may be indicative of knowledge flowing from A to B; Citations received can tell the "importance" of the cited patent. One can tell from the table that the average citations made and received are about 7 and 4, but the maximum citations made and received both reach more than 700.

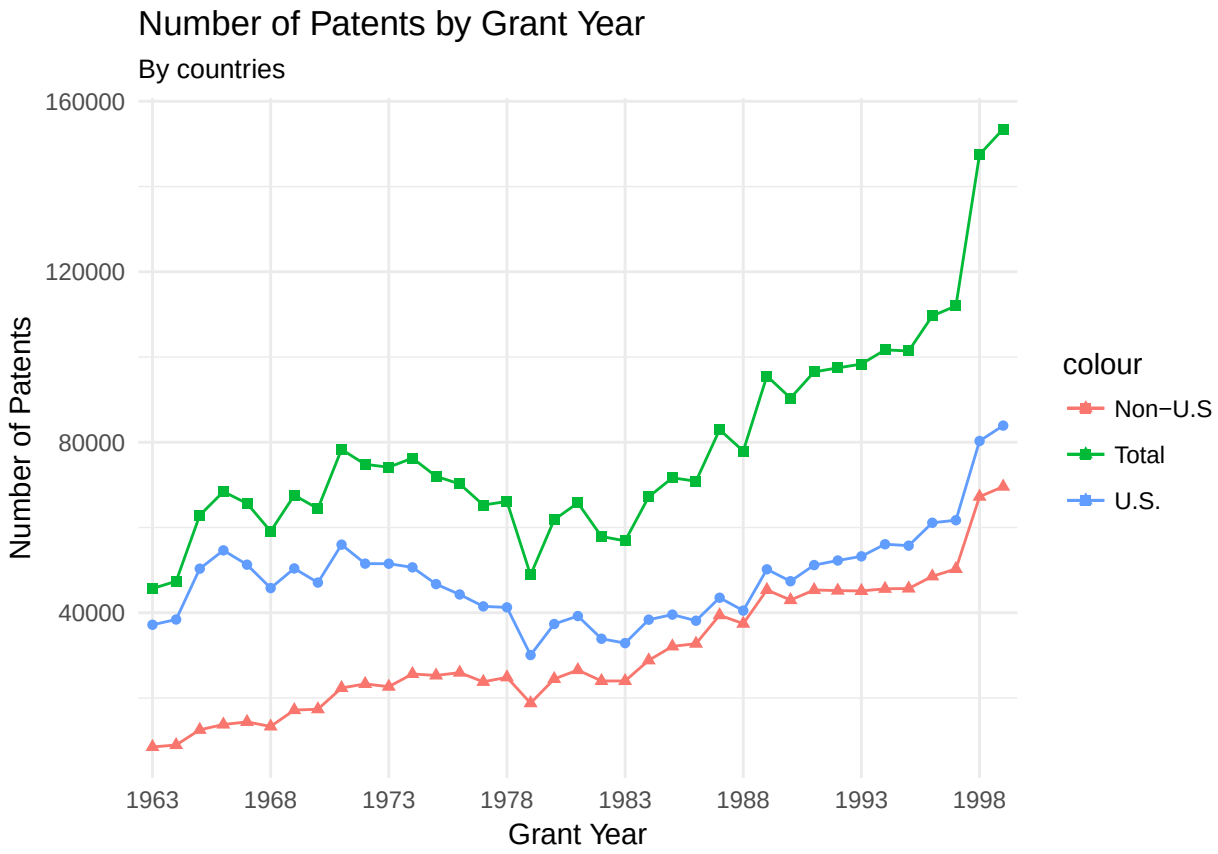
The citation lags focus on the time difference between the application or grant year of the citing patent, and that of the cited patents. Backward and forward lags look at the time difference from different directions. From the table one can see their average value are 8 and 14.

The curators of this dataset constructed their own measure of generality and originality, which range from 0 to 1. Their mean values are about 0.33.

Self-Citations focuses on whose patents are cited, and in particular, to what extent they cite previous inventions patented by the same assignee, rather than patents of other, unrelated assignees. This implies technological spillovers: presumably citations to patents that belong to the same assignee represent transfers of knowledge that are mostly internalized, whereas citations to patents of "others" are closer to the pure notion of (diffused) spillovers.

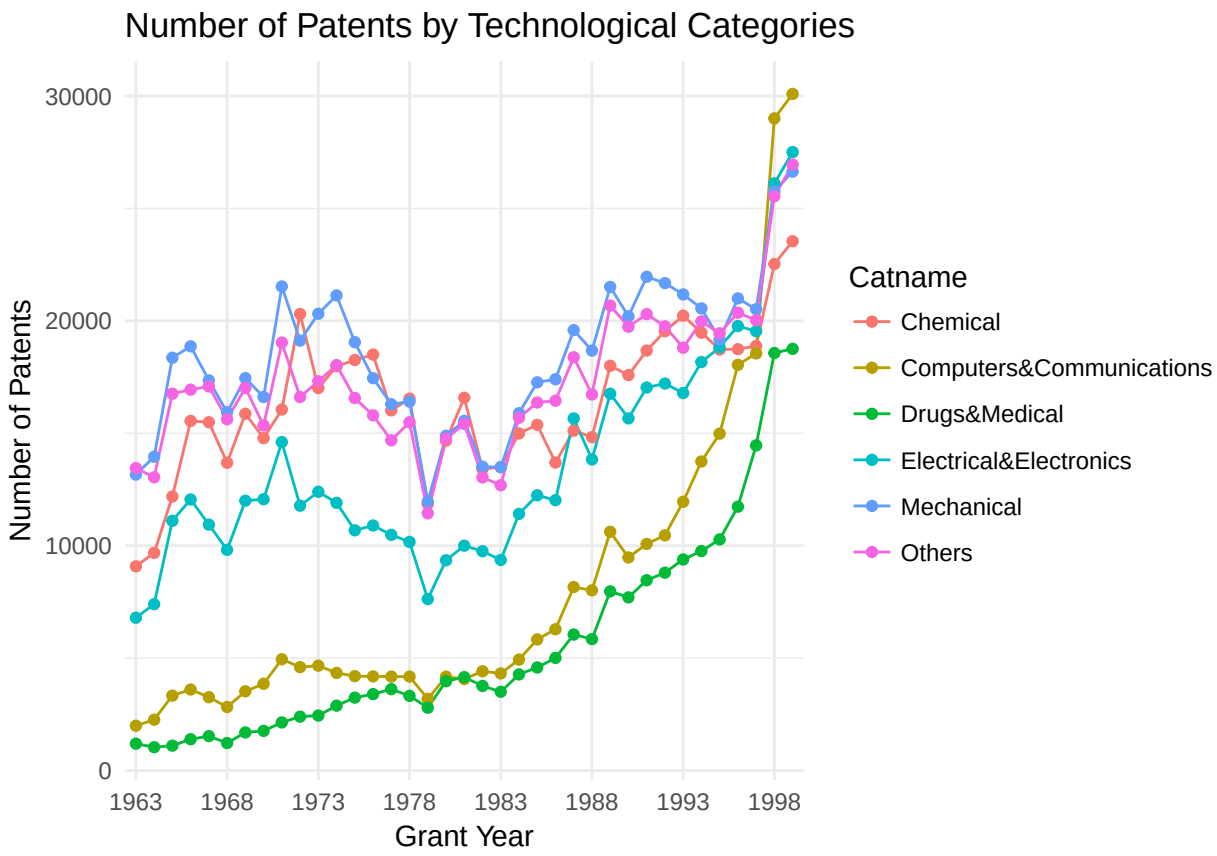
Characteristics and Conditional Description of the data

Number of Patents by Grant Year

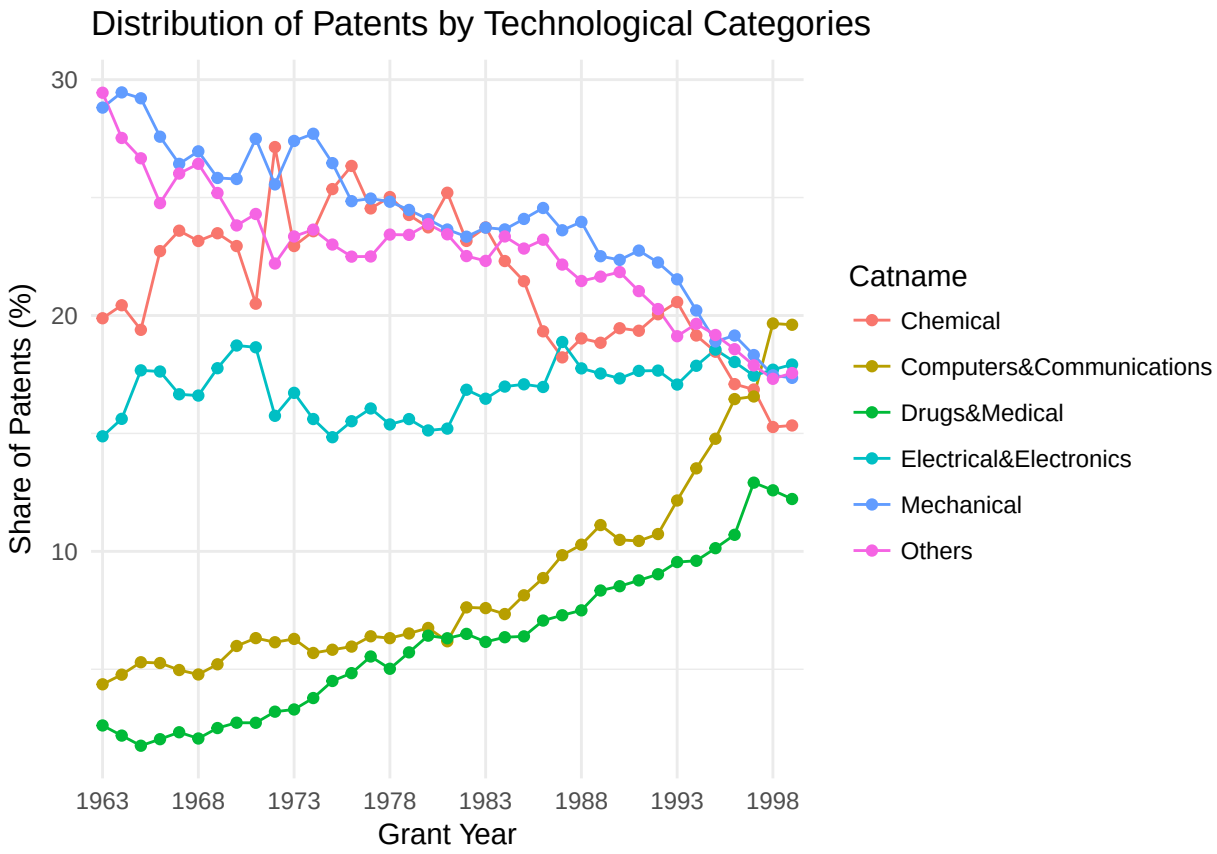


From the above graph, one can tell the number of patents is generally increasing in these 30 years. In the year 1997, there seems to have a rapid increase in both Non-US contries and the US. In general, there's more U.S. patents granted, which is reasonable for USPTO.

Distribution of Patents by Technological Categories



Looking at number of patents granted by technology categories, there are some interesting characteristcs. Patents in the field of computers & communications have increased more rapidly than patents in other fields, which surpassed the number of other patents in the year 1997. All fields have relatively more patents granted in the year 1997 than other years.



Apart from absolute number, one can also see the distribution of patents by technological categories by looking at the share of patents. From the above graph, one can tell the patents in computers & communications has been more and more important, while mechanical patents, a once predominant technological field has less importance in recent years.