Perspectives on Computational Social Science

Problem Set #1

Xingyun Wu

19/4/2017

1. Data Introduction and Access

The dataset I use is the NBER U.S. Patent Citations Data, which was collected and managed by the National Bureau of Economic Research. To be more specific, the exact data I would use is its patent data including constructed variables. There are 2,923,922 observations in this dataset, coming from all the U.S. utility patents granted between January 1963 and December 1999. Apart from basic information of the patents ("original variables", which consists of time, location of inventors, claims, and class), the dataset also contains information of their citations ("constructed variables", which consists of evaluations of each patent from different dimension). Furthermore, the collected patent covers six main technological categories: Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical, and Others. And for each technological category, there are several sub-categories.

NBER provides easy access to this dataset. This data, together with related documentations, now could be freely downloaded from its official website, under the website of the National Bureau of Economic Research.

2. Other key papers using this data

Most of the key papers having used this data are published on Research Policy, economic articles, and articles about innovation and technology. And the main concern of these papers is effects of patent policies.

Researchers have reviewed major changes in patent policy and their corresponding effects (Jaffe, 2000; Trajtenberg, 2001). More specifically, researchers have also analyzed the effect of patent policy on specific phenomenon, such as knowledge diffusion (Stolpe, 2002). But it is also indicated that problems might occur when highly interdependent technologies are attempted to be combined (Fleming & Sorenson, 2001).

3. Data Collection

The data was produced with two means. The "original variables" came from records of the U.S. Patent Office, while the "constructed variables" were generated from the original data.

Table 1: Classification of Variables	
Original Variables	Constructed Variables
Patent Number	Technological Category
Grant Year	Technological Sub-Category
Grant Date	Numer of Citations Made
Application Year	Numer of Citations Received
Country of First Inventor	Percent: Citations Made / Patents Granted)
State of First Inventor (if US)	Measure of Generality
Assignee Identifier (missing 1963-1967)	Measure of Originality
Assignee Type	Mean Forward Citation Lag
Number of Claims	Mean Backward Citation Lag
Main Patent Class (3 digit)	Share of Self-Citations Made
	Share of Self-Citations Received

4. Descriptive statistics

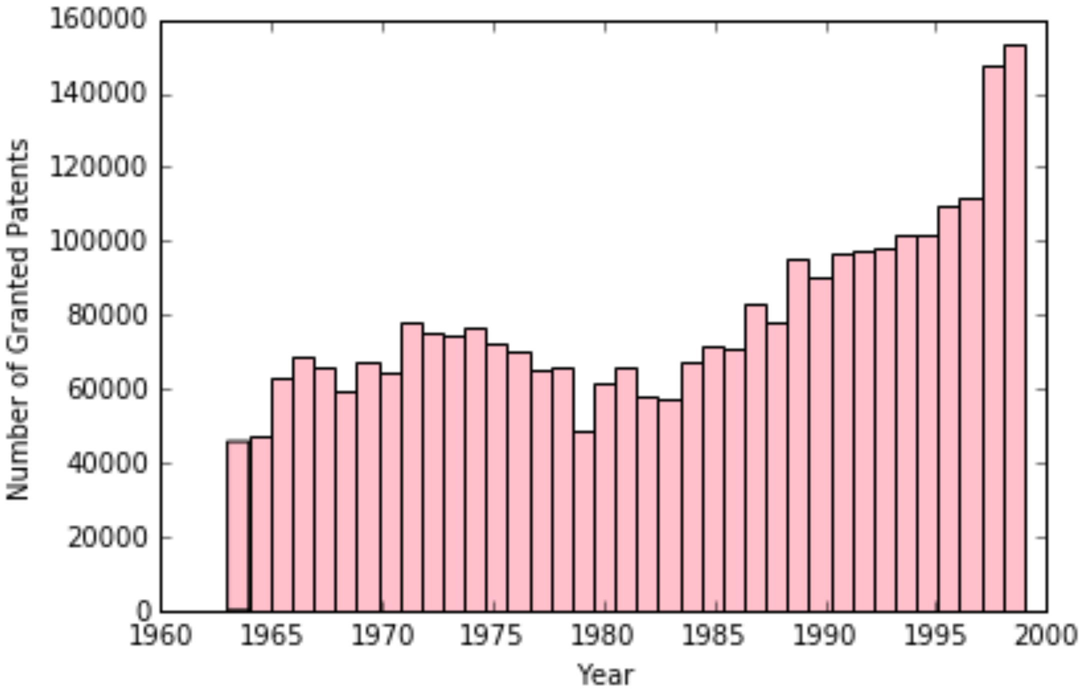
Since time, location of inventors, influence, and evaluations of generality and originality are key features of any patents, 10 corresponding variables are chosen for descriptive statistics. To make the structure more clear, I separate them into to two parts: continuous variables and categorical variables. Although variables of year is often treated as continuous variables in data analysis, it would be meaningless to treat them as continuous variables and calculate their mean and standard deviation. So in this part, I put them into categorical variables. And for categories variables with more than 10 categories, I aggregate them into fewer classes or cohorts.

Table 2: Data Description						
Continuous Variables						
Variables	Type	Min	Max	Mean	Std	Observations
Num. of Citations Made	float	0	770	7.720	9.000	2,139,314
Num. of Citations Received	float	0	779	4.779	7.346	2,923,922
Generality	float	0	0.940	0.321	0.285	2,240,348
Originality	float	0	0.951	0.349	0.281	2,042,151
Categorical Variables						
Variables	Type	Num. of Categories			Num. of Sub-Categories	Observations
Grant Year	int	37 years => 4 cohorts			37	2,923,922
		1963-1970			8	481,060
		1971-1980			10	687,818
		1981-1990			10	737,017
		1991-1999			9	1,018,027
Application Year	int	37 years => 5 cohorts			37	2,699,606
		<1963			41	6,241
		1963-1970			8	414,390
		1971-1980			10	658,086
		1981-1990			10	771,006
		1991-1999			9	849,883
Counry of First Inventor	str	162 => 2 cohorts			162	2,923,922
		US			0	1,784,989
		Non-US			161	1,138,933
State of First Inventor	str				57	1,784,989
Assignee Type	int				7	2,923,922
		unassigned				537,988
		U.S. NGO				1,380,310
		Non-U.S. NGO				913,470
		U.S. Individual				24,097
		Non-U.S. Individual				9,146
		U.S. Government				48,323
		Non-U.S. Government				10,588
Technological Category	int				6	2,923,922
		Chemical			6	606,934
		Computers & Communications			4	290,337
		Drugs & Medical			4	204,199
		Electrical & Electronics			7	499,741
		Mechanical			6	681,378
		Others			9	641,333

5. Key Visualization

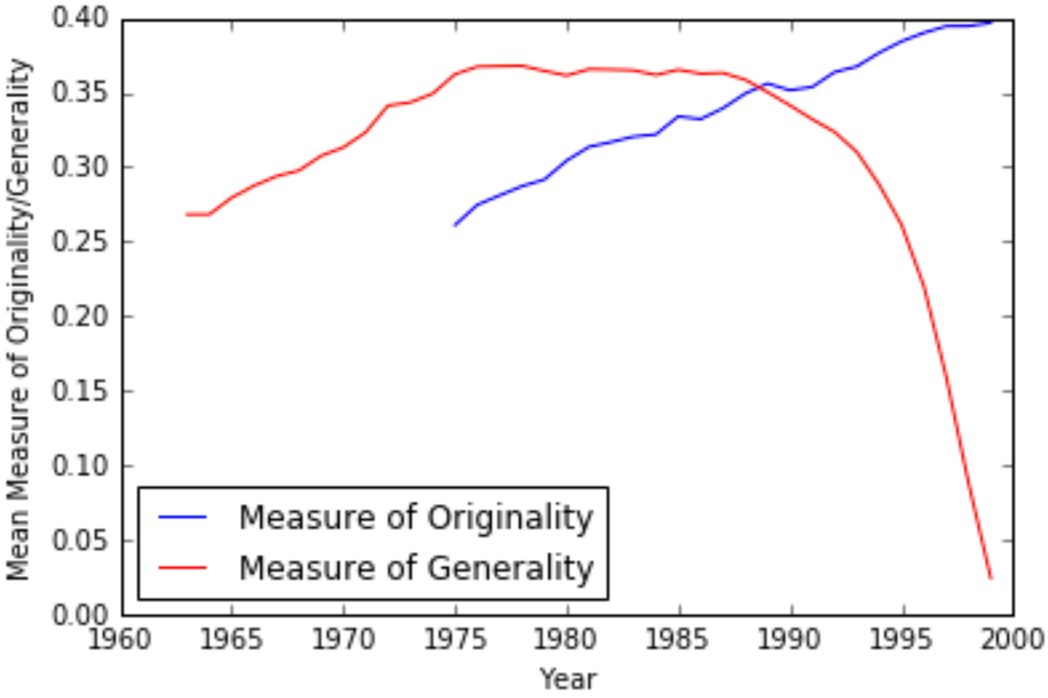
Patent production is a very significant indicator of technological innovation. Figure 1 shows number of patents produced each year. According to the histogram, from 1963 to 1983, number of granted patents fluctuated within a limited range, but the trend is not positive. In 1979, the production of patents even reached its lowest point. Since 1980, number of granted patents has been increasing. The number even significantly increased after 1997, exceeding 150,000 for 1998 and 1999.

Figure 1: Number of Granted Patents by Year



Generality and originality are also important features of patent production. In the more recent years, patterns of originality and generality have changed along the years of observation. According to Figure 2, the level of generality gradually increased from 1963 to 1975. Then it slightly fluctuated from 1975 to 1982, and has greatly decreased since 1983. In contrast, the level of originality has steadily increased since 1975. Although the data does not provide measure of originality from 1963 to 1974, the trend is still clearly shown by the plot. The comparison may mean that originality has grown to be more important than generality in recent years.

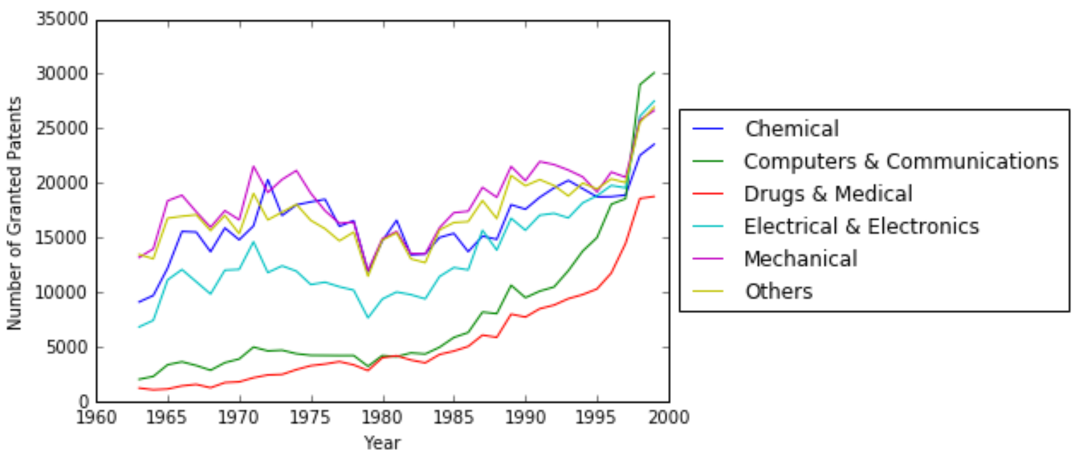
Figure 2: Mean Measure of Originality/Generality by Year



6. Conditional (slice) description

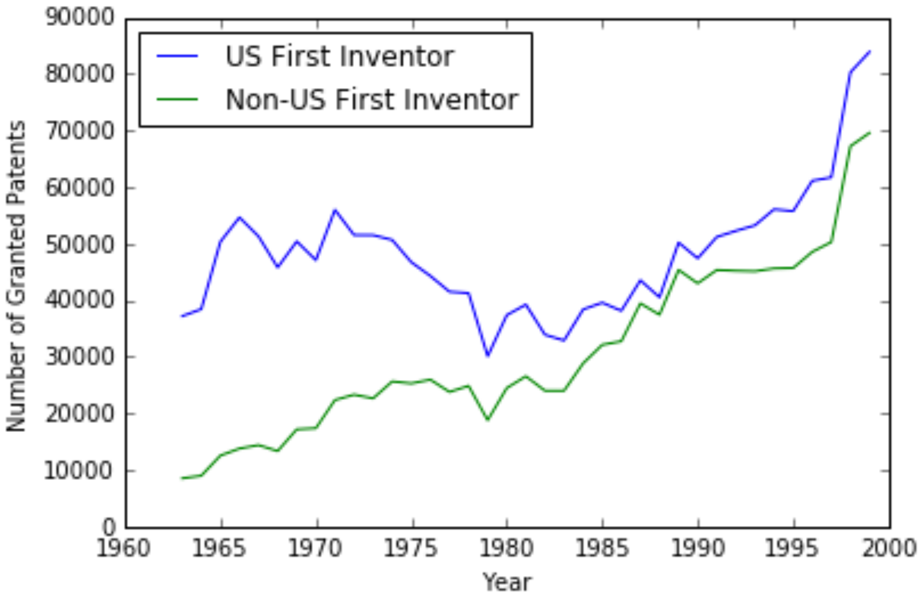
Patent production could have different patterns among disciplines. Figure 3 shows the number of patents of each year under each technological category. For Computers & Communications, and Drugs & Medical categories, the number of granted patents has kept growing. Particularly, it grew fast after 1990. But for the other four disciplines, the number of granted patents was fluctuating from 1963 to about 1997, then greatly increased in 1998 and 1999.

Figure 3: Number of Granted Patents of Each Year by Technological Category



It is also interesting to look at where the first inventors belong. According to Figure 4, at the beginning, the number of U.S. first inventors was almost 4 times as the number of non-U.S. first inventors. Then the number of non-U.S. first inventors kept growing, while the number of U.S. first inventors was fluctuating. Although the production of U.S. inventors started to increase after 1985, the difference between U.S. and non-U.S. has shrunk. In 1999, the number of U.S. first inventors is only 1.2 times of the number of non-U.S. first inventors.

Figure 4: Number of Granted Patents of Each Year by Country of First Inventor



7. References

[1] Hall B H, Jaffe A B, Trajtenberg M. The NBER patent citation data file: Lessons, insights and methodological tools[R]. National Bureau of Economic Research, 2001.

[2] Fleming L, Sorenson O. Technology as a complex adaptive system: evidence from patent data[J]. Research policy, 2001, 30(7): 1019-1039.

[3] Jaffe A B. The US patent system in transition: policy innovation and the innovation process[J]. Research policy, 2000, 29(4): 531-557.

[4] Trajtenberg M. Innovation in Israel 1968–1997: a comparative analysis using patent data[J]. Research Policy, 2001, 30(3): 363-389.

[5] Stolpe M. Determinants of knowledge diffusion as evidenced in patent data: the case of liquid crystal display technology[J]. Research Policy, 2002, 31(7): 1181-1198.