Problem set #1: By Cooper Nederhood

Part 1: Write a data section for your assigned data set (5 points)

Describe how to access data, where it is stored, who curates it. Make sure to use the original source and curator in addition to the NBER site to which I have linked

1. The data is originally described in a paper by Hall, Jaffe, and Trajtenberg and this is the original source of the data. There is a corresponding CD containing the complete dataset itself which can be ordered from MIT Press. The NBER is working on an NSF to update and extend the data. Iain Cockburn is the PI. The updated files can be downloaded at the redirect link to the updates. The data is freely available in two formats: a SAS format and a csv format. Additional more recent data can be obtained from the US Patent's ftp site.

Cite other key papers that have used this data

1. The data has been used widely and some of the papers using the data, which have the most citations are: "Market value and patent citations", (Hall et all, 2005); "Graphs over time: densification laws, shrinking diameters, and possible explanations", (Leskovec, 2005); "Patents and innovation counts as measures of regional production of new knowledge", (Anselin et al, 2002). There are many other papers citing the data but hopefully this is a sufficient amount of 'key papers'.

Describe how the data were collected.

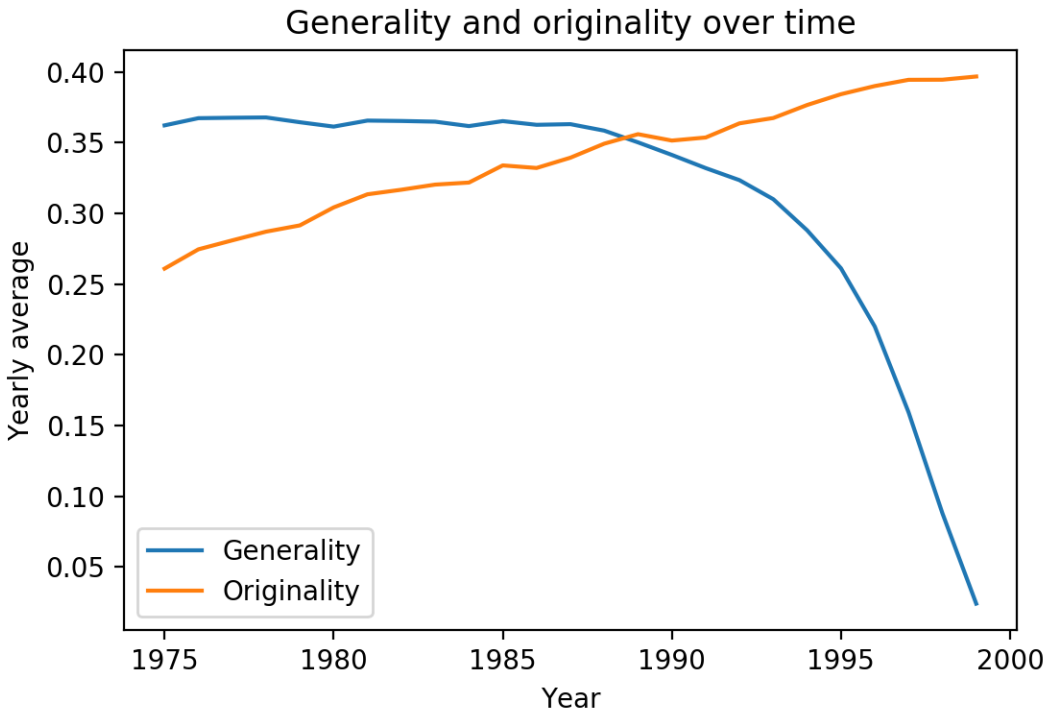
1. In the original paper the authors state that the present trends in US patenting over 30 years. In addition to the raw data, they construct a variety of original measures like backward and forward citation lags. They break the data into six main technological categories. The original data contains about 3 million patents and 16 million citations. Gathering the data was a long process and a culmination of a large array of people and projects.

Include a table that gives descriptive statistics for at least 8 key variables (you can do more).

1. See tables at end of PDF. I have included tabulations for discrete variables and summary stats for continuous variables. The discrete variable tabulations I have sorted by value. Enjoy!

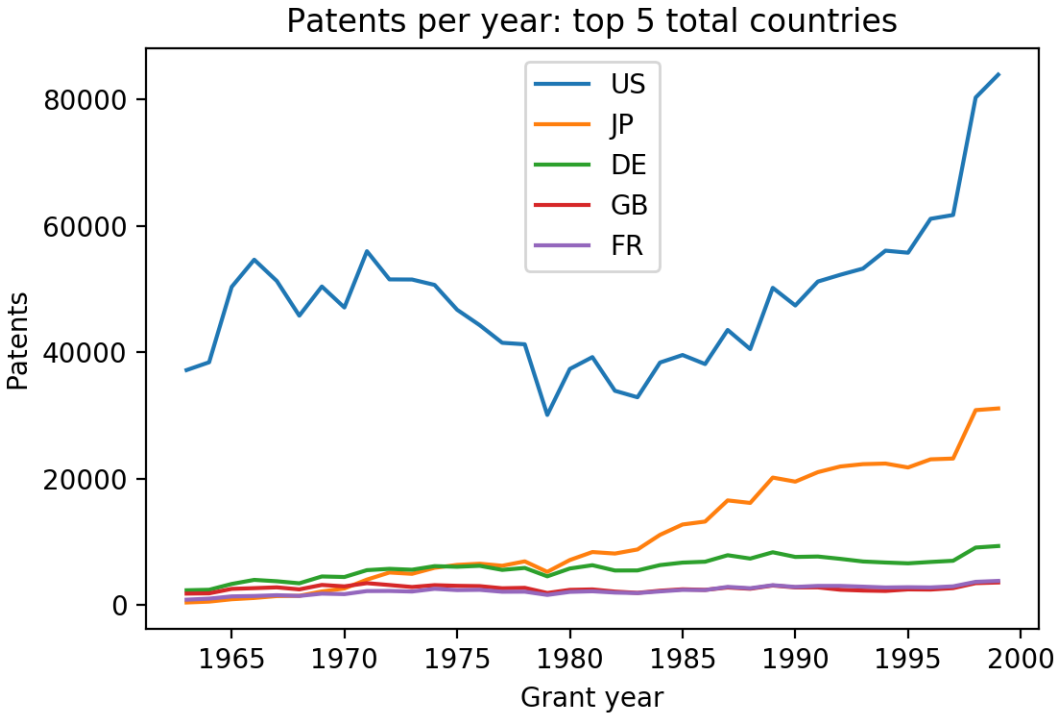
Include at least one key visualization of the data that exhibits an interesting characteristic.

1. In the chart below I have graphed generality and originality measures over time. As you can see, the generality measure has a sharp decline which is increasingly pronounced as the years progress. In contrast, the originality measure is slowly trending upwards. Both trends suggest avenues for future inquiry.


Generality and originality over time
Yearly average
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
1975
1980
1985
1990
1995
2000
Year
Generality
Originality

Show at least one conditional (slice) description of the data (e.g., all variable descriptive statistics by nationality of survey respondent). This can be a table or visualization

In the graph below I have filtered the data on the top 5 countries by total patents produced over the entire range of the data. Then, I have graphed the total patents for each of the 5 countries per year. As you can see, the US continues to be the leader in total patents and has increased its advantage compared to DE, GB, and FR. JP (Japan), which was once similar to the other non-US leaders, has had a similar trend to the US and is now solidly above the other non-US countries in per year patents.



Part 2: Critique a computational research paper (5 points)

Your assignment is to write a critical review of the main computational/empirical approach to answering the research question in your assigned paper from the list below. Your critical review has no page minimum or maximum. But it should have the following components:

State the research question of your assigned paper.

1. The researches are using a series of field experiments to evaluate the quality of advice provided by life insurance agents in India. Specifically, they are addressing whether commission-motivated agents provide quality information to uninformed consumers.

What data did the paper use?

2. The researchers use data from four closely related experiments conducted in India. They trained and sent individuals to visit life insurance agents and to seek recommendations.

The trained individuals were differentiating between two life insurance options, one which was absolutely better. By varying the initial beliefs of the insurance seekers, the researchers can measure how the agents respond to features of the customer interaction. Are the agents responsive to customer needs? Or are they responsive to the customer's potentially incorrect beliefs? Or are they responsive to their own commission incentives?

What theory did the paper reference in order to interpret the data? (Note: it is possible that the paper has no reference to theory.)

3. The main theory the paper addresses relates to the incentives of agents that are paid by commission. Opponents of commission-based payments argue that informed agents give poor advice to ill-informed consumers to drive the consumer towards high-commission options. Proponents argue that market discipline and reputation concerns lead the informed agent to provide good advice. The paper conducts an experiment to test the competing theories.

Was your assigned paper a descriptive study, an identification exercise, a numerical solution to system of equations study, or some combination of the three? (These are the three classifications we discussed in class.)

4. The researches conduct an experiment to test the above the ories. I suppose this makes it an identification exercise because they are trying to discern a relationship between quality of advice and knowledge of the parties involved, when the agents are motivated by commissions. However, the researchers are also reporting on four closely related field experiments so there is an element of the research that is a descriptive study.

What computational methods did this paper use to answer the research question? What was their result or answer to the question?

5. As discussed above, the researchers conduct experiments to answer the research question. One consistent result which held consistently is that agents recommend whole insurance products even when the customer's script suggested term insurance would be preferable. They found that advice does respond to customer and regulatory features, this could not change the overwhelming preference for whole insurance products. By varying both the insurance product which best addresses the customer's needs and by varying the initial preferences, the researchers can isolate behavior of the agents. They find that agents do respond to customer's self-reported, and sometimes wrong, beliefs, and they respond to customer needs. This is true even when the commission on the more suitable product is higher and the agent has the incentive to correct the customer's incorrect beliefs. So, the agents cater to the customers incorrect beliefs rather than correct them.

Think of yourself as an academic referee. Give two suggestions to the author(s) of your assigned paper of things the authors might do to improve their results or strengthen their evidence for the answer to the question.

6. The authors find that even when the agent could simultaneously correct a consumer's incorrect beliefs and push the consumer toward a more appropriate product which yields a higher commission for the agent, they instead continue to cater to the consumer's incorrect beliefs. As presented, this is an illogical result as both parties would be better off by changing. The authors could explore the motivations at work behind this seemingly illogical result. For example, perhaps the agent simply wants to make a sale rather than lose the sale to a competitor. The agent may think the fastest way to a sale is by catering

to currently held beliefs, rather than risk trying to change the beliefs and losing the customer. In this framework, the apparently illogical finding seems more rational. The researchers could develop this area more.

7. An issue with experiments is their high internal validity but potentially low external validity. The opening of the article refers to dynamics between consumers and agents in western countries in the context of the financial crisis, but the current experiment is a potentially vastly different circumstance. Perhaps these findings in the life-insurance market in India do not generalize broadly to all consumer-agent interactions. The researchers could improve the scope of their research by conducting a similar experiment, perhaps on a smaller scale, in another environment to show their findings are robust to different environments.

Summary statistics for relevant variables

APPYEAR	Totals
1901	3
1902	1
1903	4
1904	4
1905	1
1908	1
1909	1
1915	1
1918	1
1921	1
1922	1
1925	1
1929	1
1930	2
1931	2
1936	1
1938	4
1939	1
1940	3
1941	5
1942	5
1943	9
1944	24
1945	26
1946	28
1947	9
1948	15
1949	16
1950	25
1951	26
1952	47
1953	39
1954	68
1955	116
1956	123
1957	141
1958	216
1959	365
1960	704
1961	1169

1962	3031
1963	12612
1964	32203
1965	54912
1966	59712
1967	60073
1968	63006
1969	65889
1970	65983
1971	66397
1972	63406
1973	66351
1974	66444
1975	65888
1976	65804
1977	65978
1978	65601
1979	65726
1980	66491
1981	63910
1982	65009
1983	61563
1984	67071
1985	71442
1986	75088
1987	81458
1988	90134
1989	96077
1990	99254
1991	100016
1992	103307
1993	106848
1994	120380
1995	137661
1996	131450
1997	114881
1998	33780
1999	1560

	BCKGTLAG
25%	6
50%	10.5

75%	18.2
count	2088785
max	154
mean	14.09978471
min	0
std	11.76879594

CAT	Totals
1	606934
2	290337
3	204199
4	499741
5	681378
6	641333

	CLAIMS
25%	5
50%	10
75%	16
count	1984055
max	868
mean	12.0828344
min	1
std	10.26838479

COUNTRY	Totals
AD	6
AE	18
AG	9
AI	1
AL	1
AM	2
AN	8
AR	850
AT	10260
AU	11386
AW	2
AZ	3
BB	9
BE	10972
BG	474

BH	3
BM	41
BN	1
BO	47
BR	1165
BS	164
BY	23
BZ	1
CA	53872
CC	1
CD	2
CH	43313
CI	4
CK	1
CL	165
CM	2
CN	811
CO	172
CR	62
CS	2121
CU	45
CY	23
CZ	58
DE	221095
DK	6479
DO	23
DZ	10
EC	31
EE	4
EG	59
ES	3601
ET	1
FI	6984
FO	1
FR	85398
GB	98012
GE	3
GF	1
GH	3
GL	1
GN	1
GP	1
GR	350

GT	58
GY	4
HK	1329
HN	27
HR	52
HT	28
HU	2399
ID	125
IE	1180
IL	7378
IN	771
IQ	9
IR	65
IS	116
IT	32433
JM	31
JO	12
JP	421441
KE	32
KG	1
KN	3
KP	1
KR	14855
KW	39
KY	33
KZ	4
LB	47
LC	1
LI	529
LK	16
LR	3
LT	7
LU	649
LV	4
LY	4
MA	57
MC	174
MG	2
MH	1
ML	25
MM	4
MO	2
MQ	2

MR	3
MT	8
MU	6
MW	3
MX	1832
MY	170
NC	1
NF	6
NG	30
NI	20
NL	26687
NO	3593
NZ	1471
OM	1
PA	46
PE	103
PF	3
PG	2
PH	195
PK	22
PL	686
PT	165
PY	8
RO	361
RU	736
SA	148
SD	7
SE	28286
SG	747
SI	58
SK	11
SM	2
SN	6
SR	2
SU	6992
SV	30
SY	8
SZ	1
TC	2
TD	1
TH	97
TN	16
TR	79

TT	56
TW	19979
TZ	10
UA	77
UG	6
US	1784989
UY	46
UZ	5
VC	1
VE	530
VG	8
VN	11
VU	1
YE	2
YU	331
ZA	3051
ZM	11
ZW	48

	FWDAPLAG
25%	4
50%	7
75%	11.5
count	2074641
max	96
mean	8.306317825
min	0
std	5.804016142

	GENERAL
25%	0
50%	0.375
75%	0.56
count	2240348
max	0.9395
mean	0.320543051
min	0
std	0.284962052

GYEAR	Totals
1963	45679

1964	47375
1965	62857
1966	68405
1967	65652
1968	59104
1969	67559
1970	64429
1971	78317
1972	74810
1973	74143
1974	76278
1975	72000
1976	70226
1977	65269
1978	66102
1979	48854
1980	61819
1981	65771
1982	57888
1983	56860
1984	67200
1985	71661
1986	70860
1987	82952
1988	77924
1989	95537
1990	90364
1991	96513
1992	97444
1993	98342
1994	101676
1995	101419
1996	109645
1997	111983
1998	147519
1999	153486

POSTATE	Totals
AK	959
AL	9086
AR	3370
AS	2

AZ	21570
CA	259277
CO	24125
CT	53803
CZ	4
DC	3185
DE	15754
FL	46256
GA	17966
GU	17
HI	1684
IA	14131
ID	6528
IL	119444
IN	39302
KS	9520
KY	10767
LA	13288
MA	74973
MD	33390
ME	2956
MI	97426
MN	43019
MO	24126
MS	3129
MT	2313
NC	24635
ND	1610
NE	4699
NH	8857
NJ	124831
NM	5641
NV	3729
NY	166957
OH	101763
OK	23677
OR	16293
PA	111595
PR	696
RI	6519
SC	10947
SD	1372
TN	16273

TX	94131
US	29
UT	10079
VA	24880
VI	65
VT	3892
WA	25402
WI	37956
WV	5837
WY	1254