The Names Game: Using Inventors Patent Data in Economic Research

Manuel Trajtenberg

Tel Aviv University, NBER and CEPR

July 2004

Plan of talk

Work in progress *(not yet paper):*

- How can we use inventors data? methodological and data construction issues
- Describe the names matching problem and methodology developed to address it
- Some preliminary statistics about the (just completed) matching of whole data set.
- Pilot on Israeli inventors
- First-cut results on their mobility.

Use of Patent Data: Main Developments

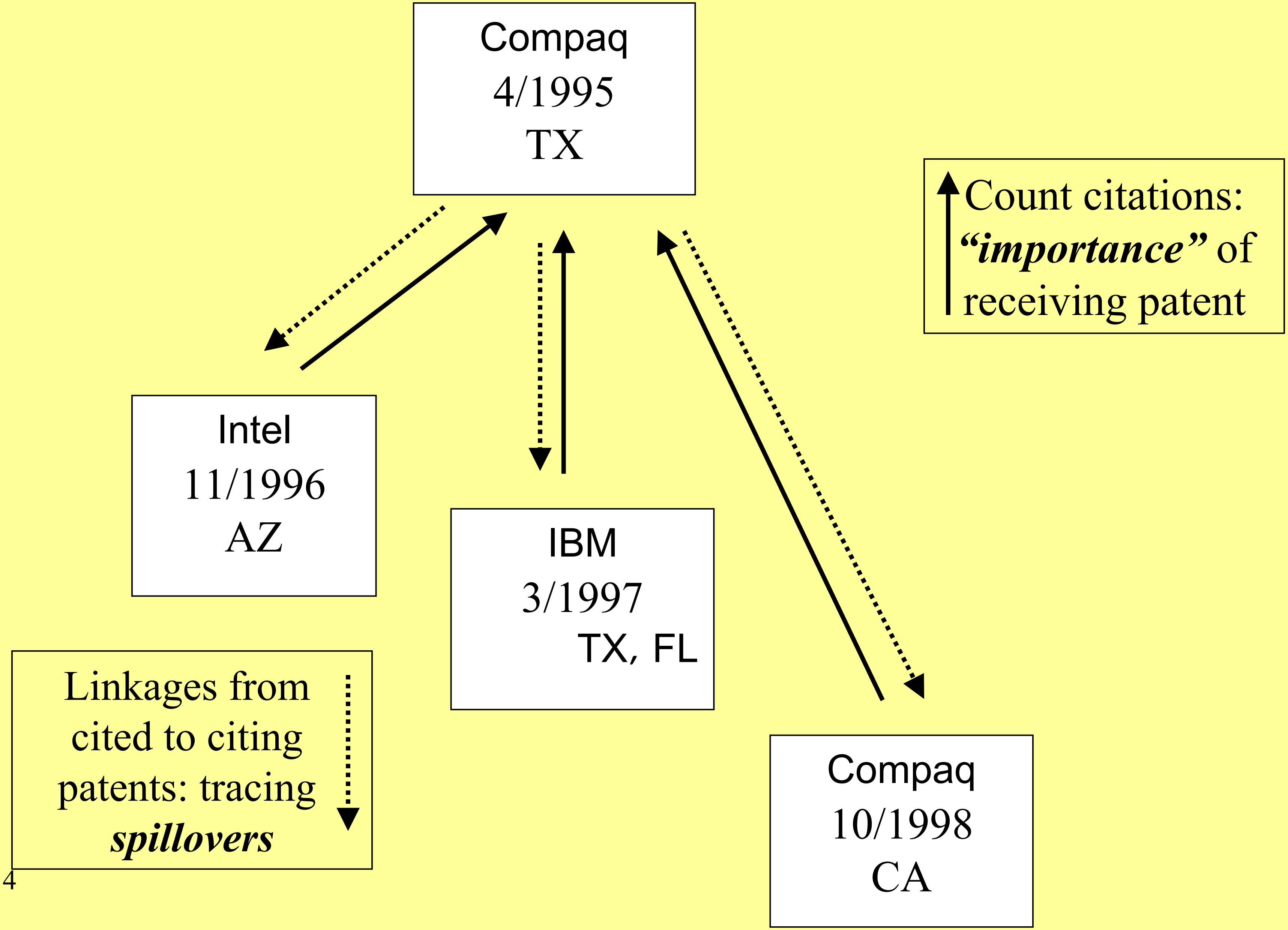
- 1960-70's: Schmookler, Scherer, etc.

- **Zvi Griliches** initiated in ~ 1980 the extensive use of computerized patent data (at the NBER); made possible the pursuit of research agenda laid out in his 1979 Rand article. Parallel use of data on patent renewals (Pakes, Schankerman).

- Early 1990's: significant step forward with the introduction of ***patent citations*** data.

- Through the 1990's: development of ***comprehensive patent & citations data*** covering ~ 30 years; late 1990's: complete data file made publicly available (NBER, J&T book).

Patent Citations: Spillovers, Importance



Compaq
4/1995
TX

Intel
11/1996
AZ

IBM
3/1997
TX, FL

Compaq
10/1998
CA

↑ Count citations: ***"importance"*** of receiving patent

Linkages from cited to citing patents: tracing ***spillovers*** ↓

Patent data used in research so far

Mostly:

- Dates (applied, granted)
- Geographical information
- Patent Tech Classification
- Assignee (e.g. linked to Compustat)
- Citations made and received
- Other: renewals, claims, litigation, etc.

Front page of patent *(partial)*

United States Patent 6,539,988

Pressurized container adapter for charging automotive systems

Inventors:

Cowan; David M. (Brooklyn, NY); **Schapers; Jochen** (New York, NY); **Trachtenberg; Saul** (New York, NY); **Nikolayev; Nikolay V.** (Flushing, NY)

Assignee: **Interdynamics, Inc.** (Brooklyn, NY)

Filed: **December 28, 2001**

Current U.S. Class:141/67; 137/614.04; 141/351; 251/149.1

Intern'l Class: B65B

Using inventors data

Vast research potential also in inventors data, not been used yet (*). Kind of research questions that could be addressed:

- spillovers through movement of inventors across countries, regions, assignees, institutions;
- “human/innovation capital” of inventors.
- productivity of R&D in firms with inventors of various characteristics;
- productivity of inventors;
- effect of work in teams and networks;
- *and more…*

The Inventors File

The NBER/Hall-Jaffe-Trajtenberg Patent Data File for 1975-1999, contains over 2 million patents, and ~ 16 million patent citations.

On average, there are about 2 inventors per patent, and thus the “Inventors File” comprises **4,298,912 records**. Each record includes *(aside from info on the patent itself):*

- The name of the inventor (Last, first, middle, surname modifier)
- Address, zip (often missing)
- City/State/Country

Who is who?

The key issue: how do we know that two records with "same/similar" names refer to the same inventor?:

1. Is Manuel Trajtenberg the same inventor as Manuel Trajtenberg?

2. Is Manuel Tra***j***tenberg the same inventor as Manuel Tra***ch***tenberg? Same as ***Em***manuel Trajtenberg?

And variants of the problem:

3. Is Manuel ***David*** Trajtenberg the same as Manuel ***D.*** Trajtenberg? As Manuel _ Trajtenberg?

Who is who – cont.

Magnitude of problem:

- Sheer ***size***: over 4 million “records” (i.e. patents x inventors)

- Have to rely ***only*** on information given in patents.

- About ½ of all patents are ***foreign*** (non-US), and hence about ½ of names non-English => idiosyncratic problems (e.g. Japanese names), what constitutes “rare/common” names, use of coding systems such as Soundex.

Work so far…

- 3- year long project – trial and error…

- Work in parallel: whole file, pilot on Israeli inventors. Learn a lot from latter, but limited usefulness because idiosyncratic, some of it cannot apply to whole file.

- Breakthrough with scoring system: allowed diagnostics, fine-tuning.

- Inherent uncertainty, but present method allows for transparent changes.

- ***Think we are done…***

Two-Stage Methodology for Matching Names

Stage 1:

- Put together records having the same (identical) inventor name (first and last, no middle for now), e.g. Manuel Trajtenberg and Manuel Trajtenberg.

- Expand the set of potential linkable names, i.e. put together Manuel Tra***j***tenberg and Manuel Tra***ch***tenberg as "suspected" of being same inventor.

"Type I error": ***if miss names that should go together; leads to under-matching, too many inventors, too little mobility, spillovers, etc.***

Methodology: second stage

Stage 2:

Link/match names deemed to be the same inventor, according to a set of criteria.

This is by far the critical and most difficult stage.

"Type II error": ***If match when shouldn't then too few inventors, too much mobility, etc.***

First stage: expand to “similar” names

Want Tra***j***tenberg and Tra***ch***tenberg to be potentially same inventor name.

Use the ***SOUNDEX*** coding method: Last name initial, followed by 3 (or more) numerical codes for consonants *(from US NARA: National Archives and Records Administration)*

Code	***Letters***
1	B F P V
2	C G J K Q S X Z
3	D T
4	L
5	M N
6	R
-	Vowels, H W Y

Soundex: examples *(using 6 digits)*

- Trajtenberg: **T623516**

(same code for *Tra***ch***tenberg*, but also for *Trestonford*…)

- Griliches: **G642200**

(same code for *Grilikes*, but also for *Garlick*…)

- Bresnahan: **B625500**

(same code for Bresnan, but also for *Brosnim*, and *Barasanam*…)

Soundex – cont.

- Clearly, expands too much! But recall that requires also same first name, e.g.: T623516_Manuel

- One way to minimize superfluous expansion: add digits – have 6 (rather than 3), but in fact 3-4 digits are enough in vast majority of cases.

- Depends upon having same last name initial (what about Yakov and Jacob).

- The system designed for English names, not well suited for e.g. oriental names, eastern European names (there exist coding systems for some of these…)

- What about first names? Could use Soundex also, but not designed for that, and does not make difference.

Second stage: stating the issue

If two records display the same name (either originally or after Soundex coding), how do we know they refer to the same inventor?

- John_Smith: 24 records
- John_ $_ Smith: 558 records
- Joh$_$_ Smith : 620 records

of which:

- John_W_Smith: 134 records
- John_W$_Smith: 141 records

The methodology of matching names

How to assess the likelihood that two records bearing the same name refer to the same inventor?

- Compare the two records according to data variables given in the patent (address, technological field, assignee, etc.); give “scores” for each matching criteria.

- Examine other possible links between them (shared “partner”, cite each other); again “scores” for them.

- Compute overall score, if above threshold then make the “match”: 120 for Soundex, 100 for identical names.

(Set threshold & scoring system considering the two types of error: over/under-matching)

Variables used for matching criteria

Name of inventor:
(Last name_first name)
Middle name *(name or initial)*
Surname Modifier *(Jr. Sr. III)*
Last name frequency
Location of inventor:
Street Address (unassigned only)
City *(size-dependent)*
State (U.S. only)
Zip code (U.S. only)
Country
more

matching criteria – cont.

Assignee *(size-dependent)*
Technological classification:
patent class *(size-dependent)*
(other?)
Citations (to each other)
Overlap of ***“partners”***

Total of ~ 10 criteria

Criteria of varying strength

- ***Strong criteria***: any one of them sufficient condition for a match, for any pair of records sharing the same Soundex-coded name.
- ***Medium criteria***: any one of them sufficient for a match of records having identical (original) names.
- ***Weak criteria***: a combination of these may be sufficient; can also support a “medium” criterion, pushing up the score so as to allow for a Soundex-based match

Strong and Medium Criteria

"Strong" criteria (120 points):

- ***Full Address***: same street address-city-country.
- ***Self Citation***: one of the records cites the other
- ***Shared partner***(s): this inventor has at least one common partner in the two records.

(implementing citations and partners: technically very complex).

"Medium" criteria (100 points):

- Same ***Middle Name***
- Same ***Zip*** (US only)

Criteria dependant upon name frequency and size thresholds

Size threshold:

The information given by the fact that two individuals are located in New York very different from the two being located in a small town. Same for assignee: two working for IBM very different from the two working for small startup.

Name frequency:

If “rare” name, then higher likelihood that two individuals with that name, plus e.g. same initial are the same guy. Not so for very common names.

Matrix of size thresholds and scores

(in terms of number of patents)

	Thresholds for Name frequency		**Score**	
	"Rare:" < 10	"Common:" ≥ 10	Below threshold	Above threshold
City	2,500	1,322 *(median)*	100	80
Assignee	2,500	500	100	80
Patent class	30,000	18,597 *(median)*	80	50

Examples of size thresholds and scores

City	**"size" of city: # of patents**	**Scores**	
		John Smith ("common")	Zvi Griliches ("rare")
Sacramento	1217	100	100
Memphis	2097	80	100
Los Altos	5968	80	80

City threshold for rare names: 2,500

City threshold for common names: 1,322

Impose Transitivity

A matched to ***B***

→

B matched to ***C***,

A matched to ***C***

Even though ***A*** and ***C*** may have little or nothing in common, except of course for (at least) same Soundex-coded name

Matching names: recap technical procedure

1. All records having the same Soundex-coded names are grouped together.
2. Each pair is examined in terms of the said criteria, and a yes-no decision to match is made on the basis of the total pair-wise score. This is done in one iteration.
3. An iterative process imposes transitivity, until convergence – complexity increases rapidly with number of records. All records matched given same ID.

An example

Inventor Name	**Partners**	**Middle name**	**City**	**Pairwise scores**	**final ID**
1. Manuel Tra***ch***tenberg	Tim Bresnahan		Boston	1-2: 120 1-3: 80	11
2. Manuel Tra***j***tenberg	Tim Bresnahan	David	Tel Aviv	2-3: 100	11
3. Manuel Tra***j***tenberg		David	Boston		11

Average matching score: 300/3=100

Diagnostics: ex post average matching score

Diagnostic tools critical: otherwise too large a file to assess the “quality” of the matches done (“manual” pilot for Israeli inventors).

Compute average matching score for each “group” of matched inventors:

- for each pair (permutation) compute the actual matching score (e.g. the sum of the points of each common criteria); there are ***m=n (n-1)/2*** permutations.

- Compute the average as: $\frac{\sum_i^m pairwise\ score_i}{m}$

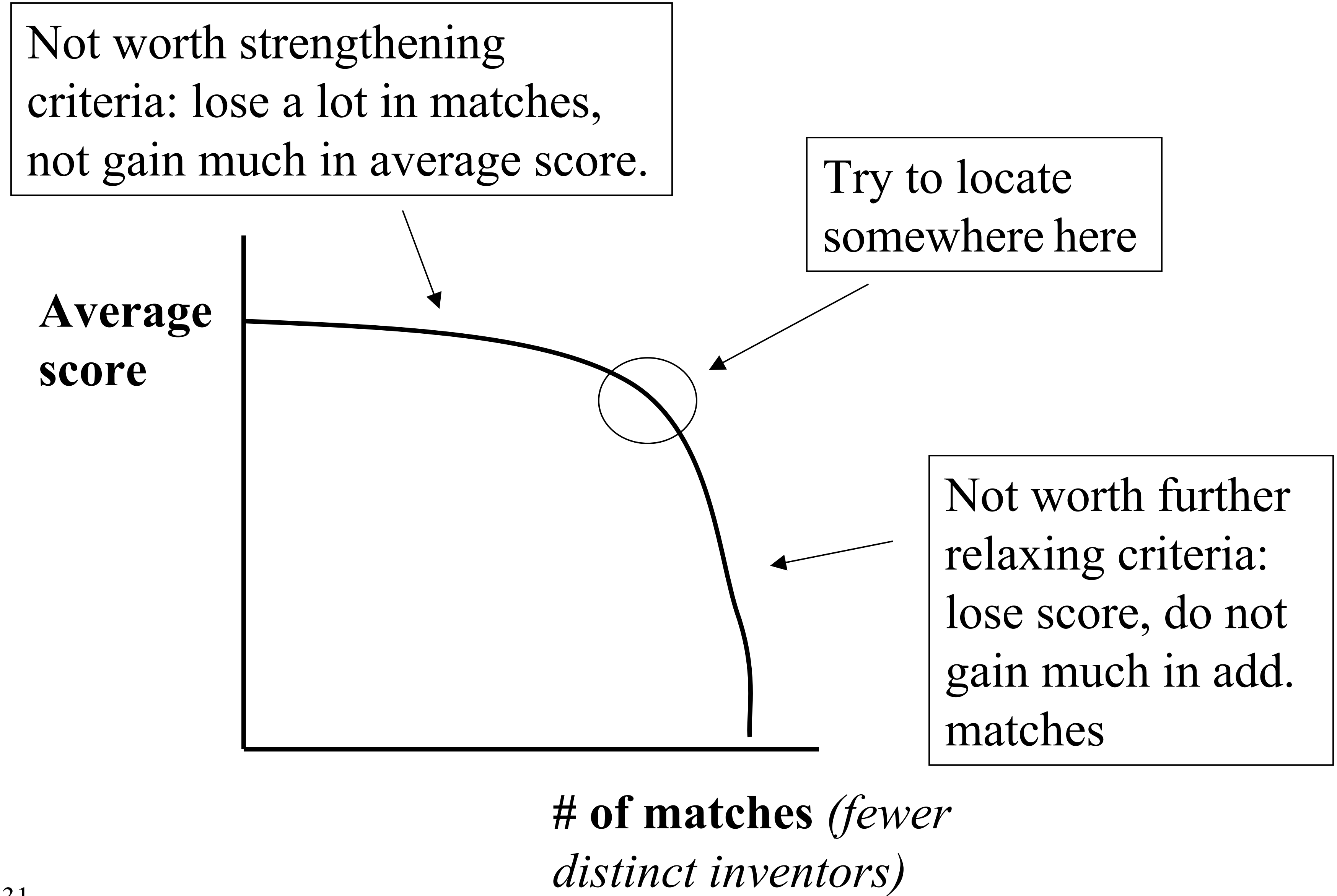
More on the average matching score

Allowed us to ***fine-tune*** the matching criteria (i.e. could define a loss function, responding to small changes in criteria).

The scores may serve as ***"weights"*** in e.g. regression analysis: give more weight to groups that their match is more certain.

The actual average matching score for the full file: ~ 240 => 2 strong criteria, or 2 medium + one weak criteria, on average among all pairs (recall transitivity…)

Trade offs between score and matches



Not worth strengthening criteria: lose a lot in matches, not gain much in average score.

Try to locate somewhere here

Not worth further relaxing criteria: lose score, do not gain much in add. matches

Average score

# of matches *(fewer distinct inventors)*

The numbers…

Original patent file:

- 2,139,313 patents
- average number of inventors per patent: 2.009

→

- 4,298,912 “records” (*patents* x *inventors*)

End result:

Matching rendered 1,565,780 distinct inventors

- Average number of patents per inventor: 2.74

Matching in perspective

No matching (each appearance of a name in a patent regarded as a different inventor):

4,300,000 *(4,298,912)*

Matching with our procedure:

1,600,000 *(1,565,780)*

“Naïve” matching - each exact [family name_ first name] a different inventor:

1,200,000 *(1,211,292)*

Naïve matching with Soundex-coded names:

800,000 *(844,171)*

Matching in perspective – cont.

The naïve 1.2 million not necessarily a subset of the 1.6 million (e.g. because of Soundex).

Huge indivisibility: either go all the way and do it all, or don't do it at all…

And now, Some summary statistics

Number of patents per inventor

(or how much "action" can we expect?)

Out of 1,565,780 inventors, the number of inventors with,

- just one patent: 911,943 (58%)
- 2 or more: 653,837 (42%)
- 5 or more: 203,302 (13%)
- 10 or more: 73,072 (5%)

Mobility of inventors across countries

Number of countries	**Number of inventors with patents>1**
1	641,127*
2	12,371
3	323
4	15
6	1
Total:	653,837
# of movers	**12,710 (1.9%)**

*Another 911,943 inventors had only one patent each, and hence could be located just in one country

Mobility of inventors across assignees

Number of assignees	Number of inventors with patents>1
1	437,256
2	158,737
3	38,727
4	11,838
5+	7,279
Total:	653,838
# of movers	**216,581 (33%)***
* But probably overstates moves: need to consolidate assignee codes.	

Mobility of inventors across US states

Number of states	**Number of US inventors with patents>1**
1	292,333
2	39,123
3	4,334
4	556
5+	120
Total:	336,466
# of movers	**44,133 (13%)**

Distribution of patents and inventors across major countries

Country	Number of Patents*	Number of Inventors**	% of Inventors
US	1,210,486	772,774	49.35
Japan	393,901	330,854	21.13
Germany	175,767	129,945	8.30
France	67,922	56,815	3.63
UK	69,375	53,570	3.42
Canada	44,767	38,237	2.44

Flows of Inventors across countries

("brain drain", "brain gain")

From

To

	US	JP	DE	FR	GB	CA	IT	CH	SE	Other	Total
US	0	808	657	265	1602	1096	68	177	113	2468	7272
JP	908	0	115	22	49	21	2	12	7	108	1244
DE	731	122	0	95	38	16	38	234	7	420	1701
FR	329	20	83	0	48	13	18	53	5	96	665
GB	2077	41	51	66	0	131	17	36	7	383	2809
CA	1308	23	11	5	106	0	5	10	7	79	1554
IT	54	2	30	17	12	4	0	37	2	28	186
CH	167	16	237	58	31	10	29	0	51	94	693
SE	164	10	12	11	11	12	3	51	0	64	338
Other	2303	72	355	126	284	89	25	92	62		4,307
Total	8041	1114	1551	665	2181	1392	205	702	261	4,657	20,769
NET	769	-130	-150	0	-628	-162	19	9	-77	350	

Flows of Inventors across US states

	NY	NJ	CA	PA	MA	CT	TX	IL	OH	Other	Total
NY	0	795	809	399	353	447	353	184	279	2,450	6069
NJ	594	0	552	599	266	231	273	187	151	1,661	4514
CA	517	360	0	323	377	199	777	333	267	4,317	7470
PA	312	483	457	0	175	107	199	185	248	1,868	4034
MA	267	190	539	175	0	153	145	114	111	1,536	3230
CT	304	185	280	123	188	0	113	103	98	838	2232
TX	199	142	745	143	108	89	0	159	166	1,897	3648
IL	167	199	530	165	128	103	219	0	198	2,112	3821
OH	256	151	357	246	121	95	236	192	0	2,112	3766
Other	1456	1040	3774	1552	1060	606	2307	1439	1465		29,227
Total	4072	3545	8043	3725	2776	2030	4622	2896	2983	33,319	68,011
NET	-1997	-969	***573***	-309	-454	-202	***974***	-925	-783	4,092	

Flows of Inventors across type of assignees

To

From

	COR	IND	GOV	Total
COR	298472	57698	5379	361549
IND	59487	0	1799	61286
GOV	7710	2024	1834	11568
Total	365669	59722	9012	434403
Net	4120	-1564	-2556	

Silicon Valley inventors
(fresh from the oven…)

44,805 inventors "related" to SV (~6% of US inventors), involved in 160,000 patents.

- 3.6 patents per inventor (>> overall mean of 2.7)
- % of assignee movers: 45% >> all inventors: 33%
- % of state movers: 16% >> all inventors: 7%
- % of country movers: 3.7% >> all inventors: 1.9%

(all percentages out of inventors with > 1 patent)

Pilot: Israeli Inventors

- Learning by doing, create benchmark, against which to assess the performance of the (computerized) matching methodology.

- Did it for all US patents granted to Israeli inventors, expanded to include all patents granted to inventors that ever had an Israeli address.

- Semi "manual" process – rendered list of unique inventors, with ***all*** their patents.

Israeli inventors: some descriptive statistics

- 6,029 Inventors, 15,316 records

- ~ 9% of inventors female *(but margin of error)*

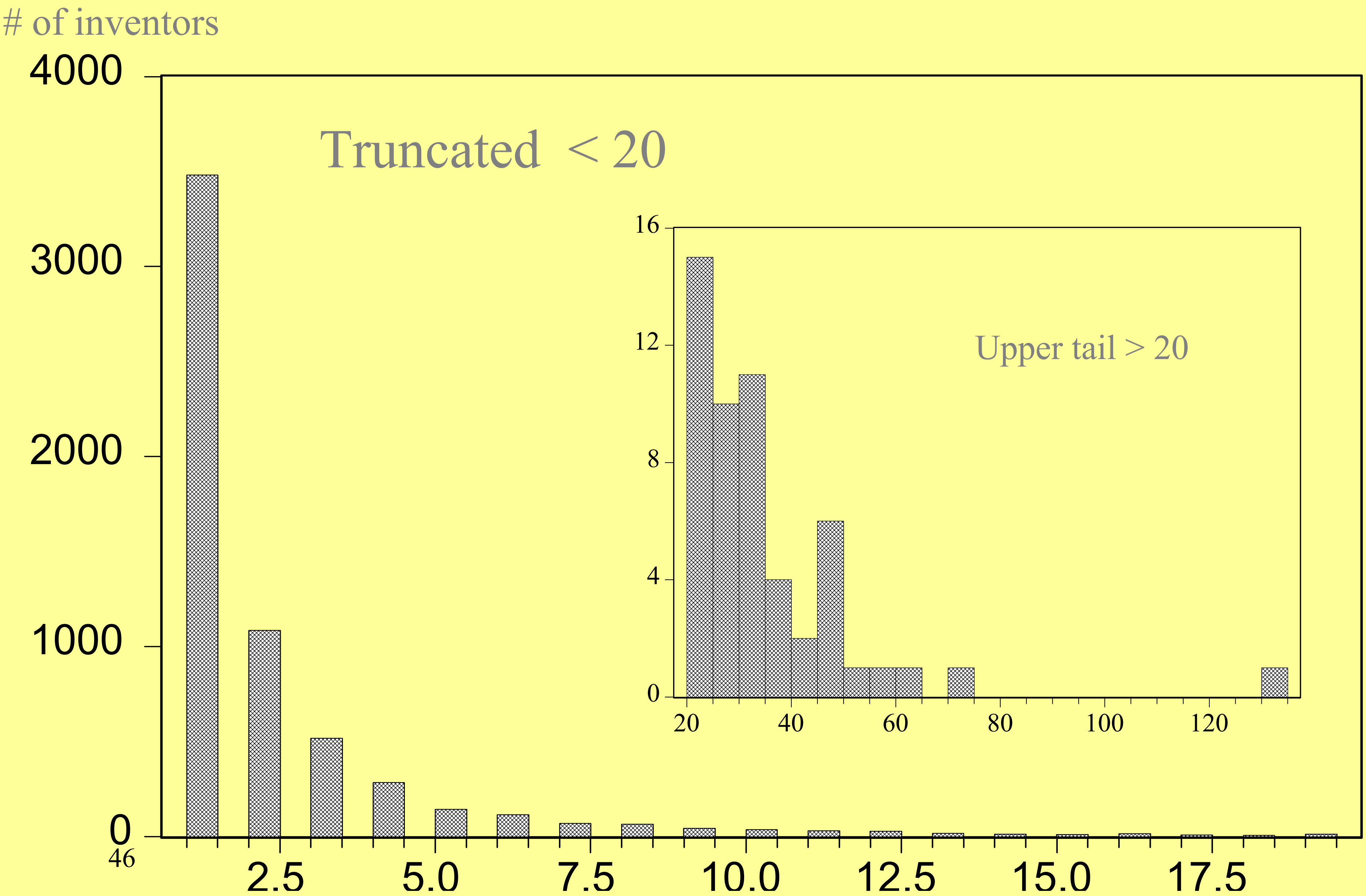
Mobility:

- 22% moved between assignees
- 6.6% moved countries *(in either direction)*

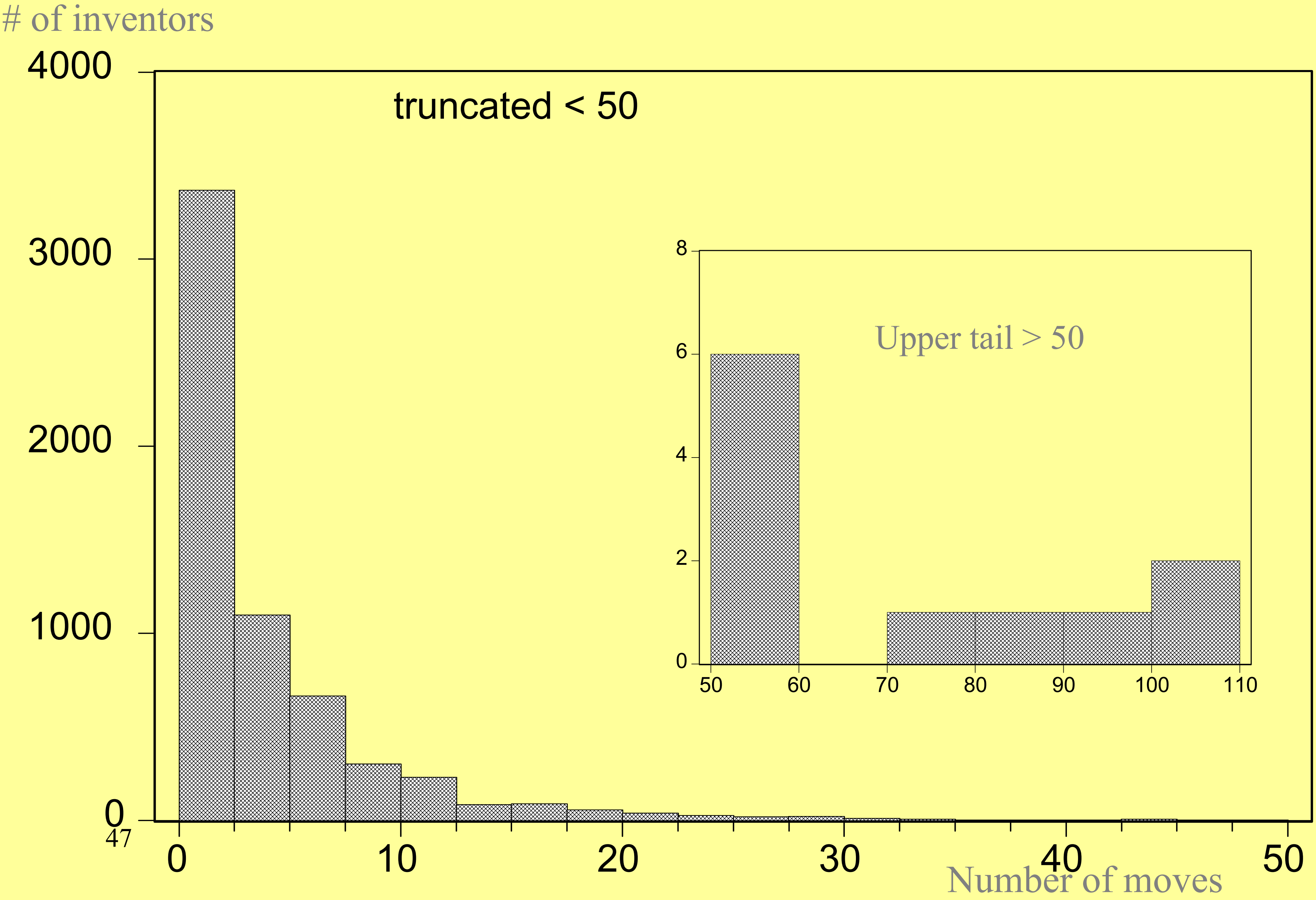
Location:

- 39% of inventors in metropolitan Tel Aviv
- 11% in Jerusalem

Number of patents per inventor

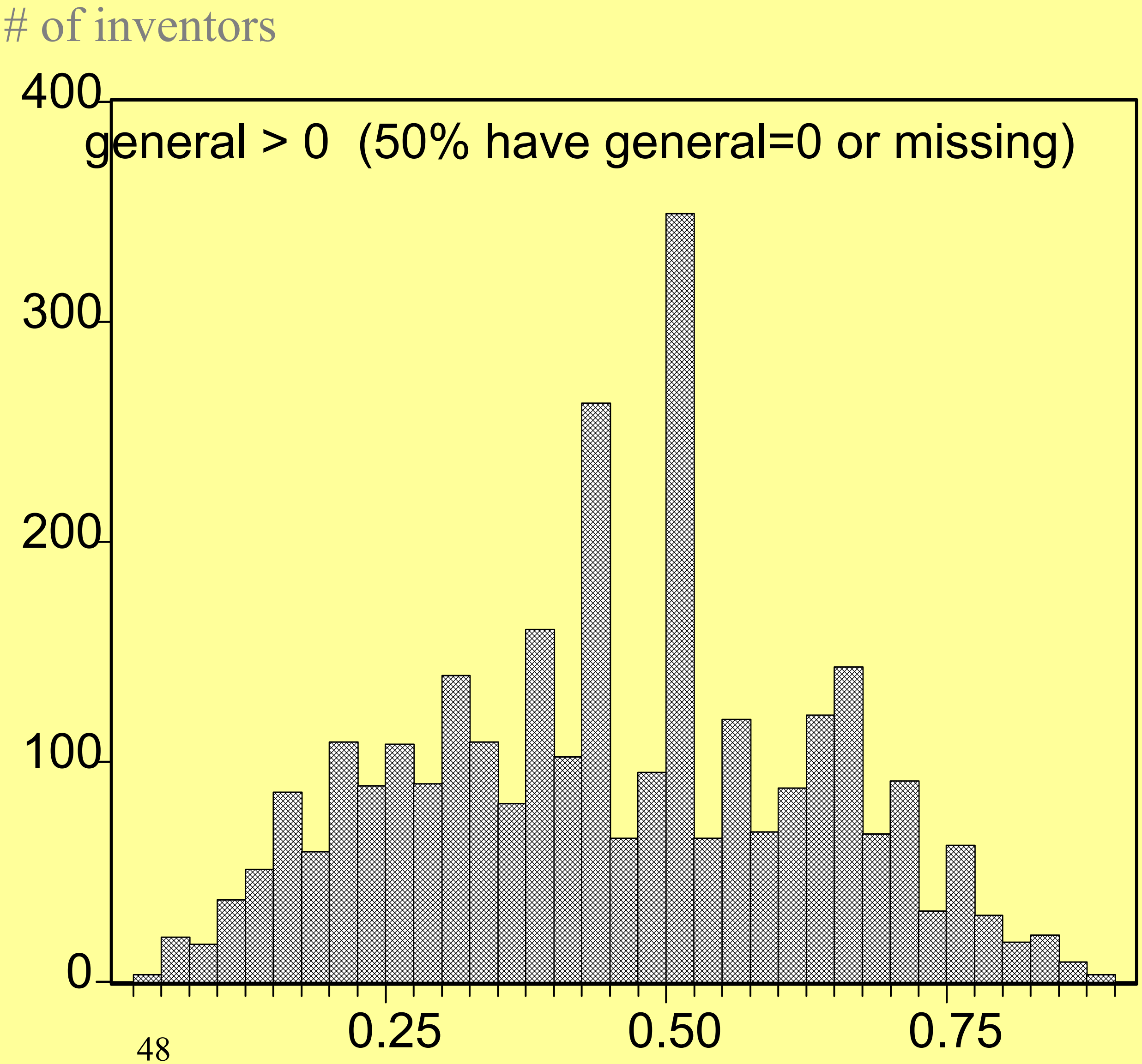


Mean citations received per inventor



Mean “generality” per inventor

(for generality>0)



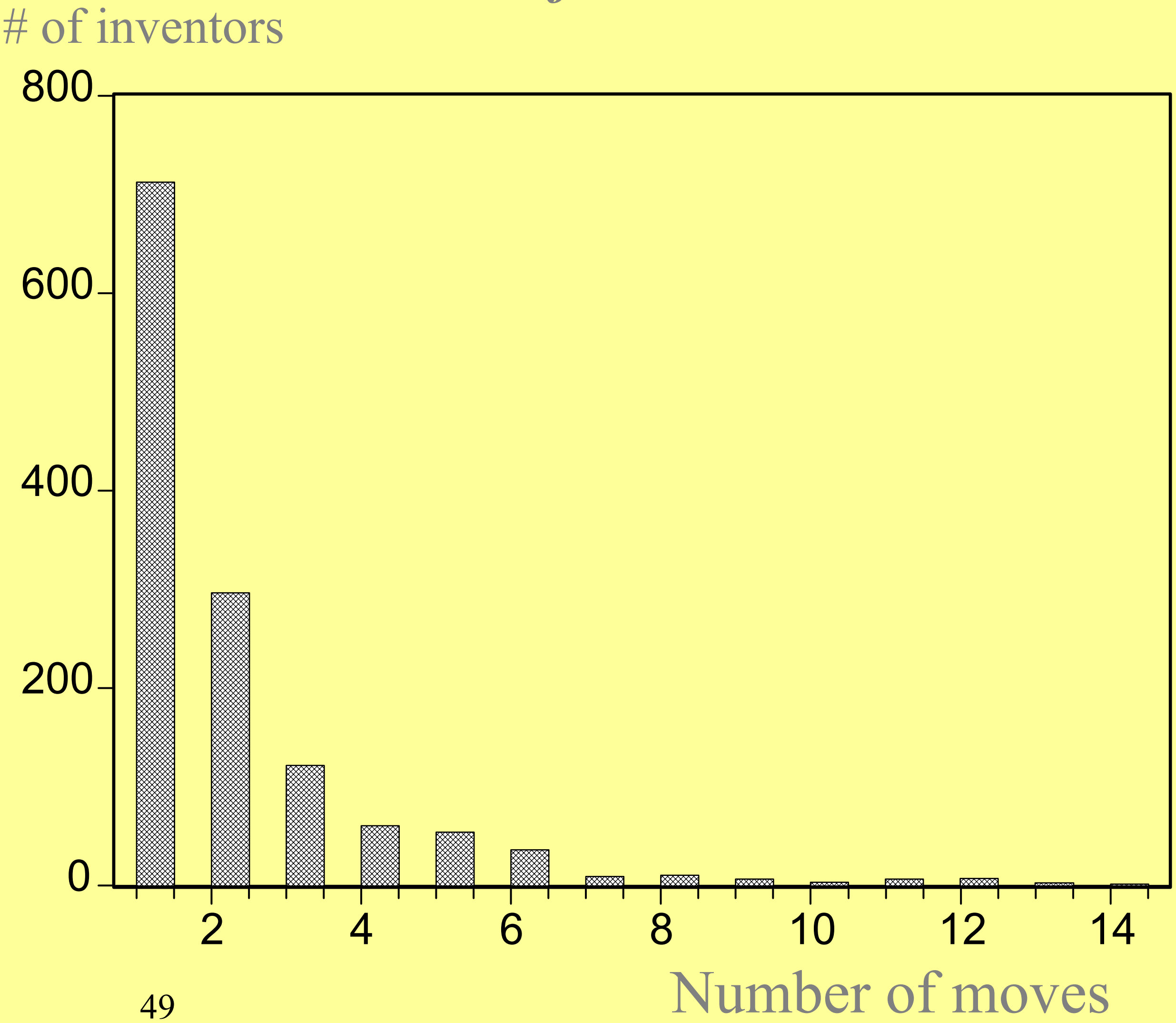
```
Series: M_GENERAL
Sample 3 6026
Observations 2969

Mean        0.448866
Median      0.444444
Maximum     0.876033
Minimum     0.036000
Std. Dev.   0.178488
Skewness    0.008342
Kurtosis    2.280072

Jarque-Bera 64.15202
Probability 0.000000
```

Number of moves between assignees per inventor

(for movers, truncated < 15)

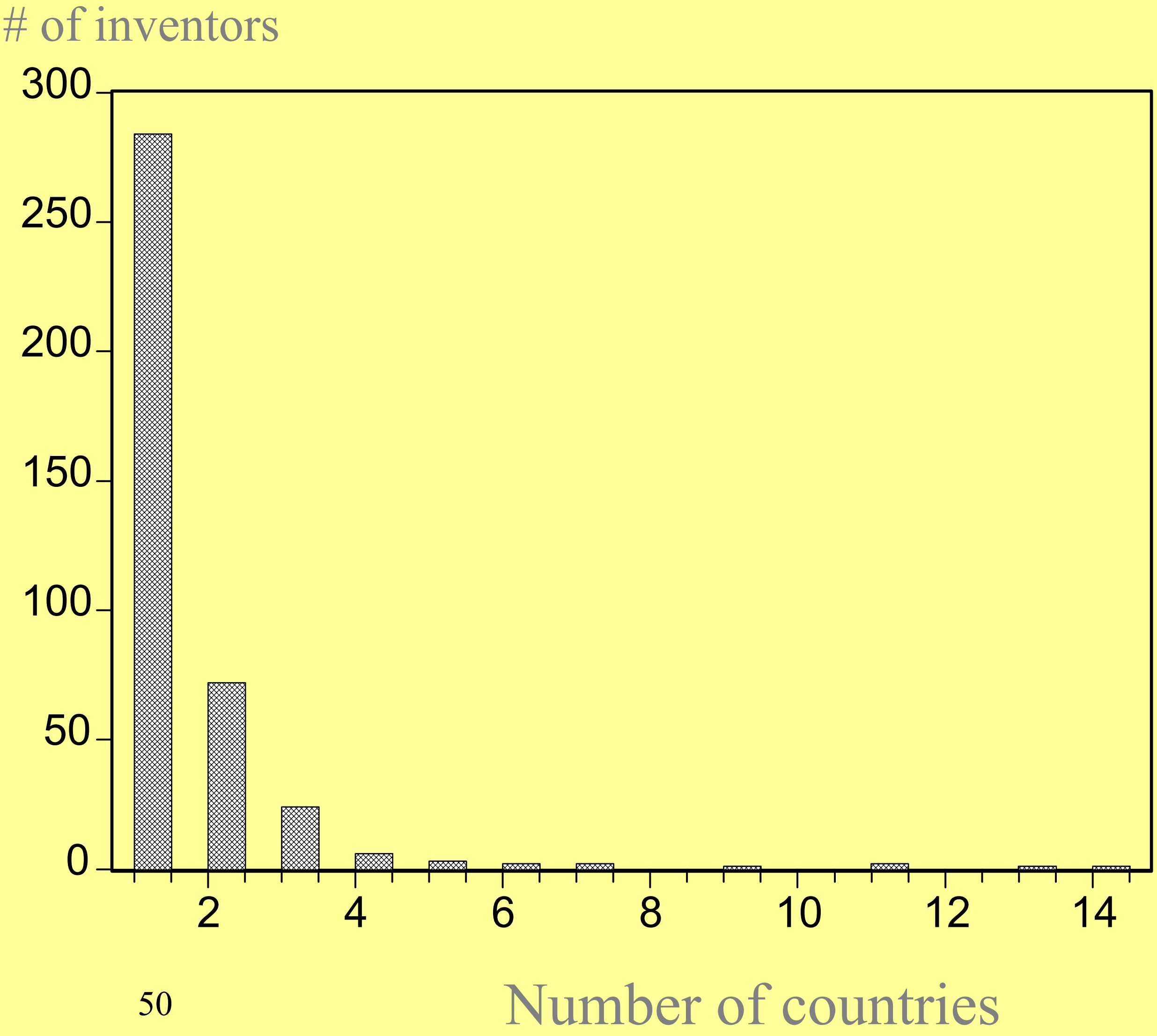


```
Series: moves assig.
Sample 2 6023
Observations 1323

Mean          2.123961
Median        1.000000
Maximum       14.00000
Minimum       1.000000
Std. Dev.     1.897062
Skewness      2.717988
Kurtosis      12.12496

Jarque-Bera   6218.908
Probability   0.000000
```

Number of moves between countries per inventor (for movers)



```
Series: N_O_G_M
Sample 26 6019
Observations 398

Mean          1.565327
Median        1.000000
Maximum       14.00000
Minimum       1.000000
Std. Dev.     1.456162
Skewness      5.094851
Kurtosis      35.48793

Jarque-Bera   19224.99
Probability   0.000000
```

Who moves between countries?

Dep. var.: no. of moves – Negative Binomial Count
Includes constant, Tech. Dummies, 6,029 obs.

	coefficient	Z-Statistic
#of patents	0.15	*10.97*
mean cites received	0.03	*5.72*
mean # of partners	-0.09	*-2.29*
% of corp. patents	0.19	*1.32*
female	-0.76	*-2.83*
LR index - pseudo R^2	0.21	

Who moves between assignees?

Dep. var.: no. of moves – Negative Binomial Count

Includes constant, Tech. Dummies, 6,029 obs.

	coefficient	Z-Statistic
#of patents	0.25	*16.64*
mean cites received	0.02	*4.86*
mean # of partners	-0.015	*-0.91*
% of corp. patents	0.19	*3.12*
female	-0.22	*-2.11*
LR index - pseudo R^2	0.25	

Who tends to move more frequently?

Both across countries and between assignees

Inventors,

- with more patents ***(but…)***
- with more “important” patents (highly cited)
- with fewer partners
- male inventors

But endogeneity!

Mobility of inventors and innovative performance

Look at “quality” of patents, as function of mobility of inventors, and controls. Dependent variables:

- Number of Citations received
- “Generality” *(1 – Herfindhal on pat classes of citing patents)*
- “Originality” *(1 – Herfindhal on pat classes of cited patents)*
- Number of Claims

Dep. variable: citations received

OLS, 15,316 obs (patents), include constant, dummies for tech field, and for assignee type

	1	2	3
Grant Year	-0.47 *(-36)*		
Patent seq. of inventor	-0.01 *(-1.4)*		
# of partners	0.13 *(4.2)*		
Moved countries	**1.37** ***(6.0)***		**1.5** ***(5.7)***
# of former country moves		**0.16** ***(2.6)***	**-0.1** ***(-1.4)***
# of former assignee moves			**0.01** ***(0.6)***
R2	0.15	0.15	0.15

Other Indicators of Patent "Quality"

OLS, 15,316 obs (patents), include constant, dummies for tech field, and for assignee type

	Generality	**Originality**	**Claims**
Grant Year	-0.01 *(-22)*	0.007 *(17)*	0.27 *(14.8)*
Patent seq. of inventor	-0.001 *(-4.1)*	-0.001 *(-2.9)*	0.02 *(1.1)*
# of partners	0.008 *(5.5)*	0.01 *(11.9)*	0.34 *(3.6)*
Move countries	**0.40** ***(4.4)***	**0.02** ***(3.0)***	**1.51** ***(3.7)***
# of former geo moves	**0.009** ***(2.3)***	**0.01** ***(4.1)***	**0.18** ***(1.0)***
# of former assig. moves	**0.0005** ***(0.4)***	**0.002** ***(2.1)***	**0.19** ***(3.2)***
R2	0.074	0.056	0.055

Mobility – Main Findings

- Inventors that move have on average more and better patents, but ***simultaneity:***

- Moving impacts favorably the quality of patents

- Moving countries has the largest effect, moving between assignees less so.

- The effect seems to come immediately, past moves have a lesser impact.

- More partners decrease the probability of moving, but increase the quality of patents.

Further work

- Study impact of inventors' mobility on firms' innovative performance, *both ways!*

- Use together both data on mobility of inventors and on citations to trace spillovers

- Study mobility of inventors between regions and firms, as function of regional and firm-related variables.

- etc….