On-line appendixes to “PROGRESS AND POTENTIAL: A profile of women inventors on U.S. patents”

Published February 2019

Andrew A. Toole
Office of the Chief Economist, USPTO

Amanda Myers
Office of the Chief Economist, USPTO

Stefano Breschi
ICRIOS—Università Bocconi,

Edoardo Ferrucci
LUISS Business School

Francesco Lissoni (contact author – francesco.lissoni@u-bordeaux.fr)
GREThA UMR CNRS 5113—Université de Bordeaux

Ernest Miguelez (contact author – ernest.miguelez@u-bordeaux.fr) GREThA UMR CNRS 5113—Université de Bordeaux

Valerio Sterzi
GREThA UMR CNRS 5113—Universite´ de Bordeaux

Gianluca Tarasconi
ICRIOS—Università Bocconi,

Table of contents

Appendix 1: Methodology

This appendix describes the methodology used to attribute gender to USPTO PatentsView disambiguated inventors. It begins with a review of the literature on gender attribution (A.1.1.) and proceeds to a detailed description of the attribution methodology (A.1.2) that was applied to the names of inventors listed on granted U.S. patents from 1976 through December 2016. The method leverages on two main data sources – IBM's Global Name Recognition dataset and the WIPO Worldwide Gender-Name Dictionary.

A.1.1. Literature review

The vast majority of attempts to attribute gender to patent inventors use methods that compare inventors' names with lists of worldwide names that suggest a feminine or masculine characterization.[1]

The U.S. Patent and Trademark Office (USPTO) has a long history of studies on women inventors. The first report, called "Buttons to Biotech -- U.S. Patenting By Women, 1977 to 1988," was published in January of 1990. This report was updated twice, once in 1998 and again in 1999. In all of these early reports, the USPTO matched given names of patent inventors as listed on the front page of each patent to a file of female-only given names. Inventors having given names that are male-only or could not be easily characterized as male or female were assumed to be male.

Naldi et al. (2004a, 2004b) is one of the first systematic attempts in the academic literature to classify inventors by gender. The authors collected 8,291 unique names from such sources as dictionaries, calendars, books and internet sites, files from "Record Offices," and phone books for six European countries. They used that list to compare approximately 100,000 unique names of inventors residing in those same six countries who also appeared in published European Patent Office (EPO) patents in 1998. In their results, 97.2% of inventors' names where attributed a gender. Their approach, though country specific (and addresses the 'Andrea' problem – name gender attribution varying with language), does not account for the potential migratory background of inventors (a Spanish-resident, Italian-origin inventor named 'Andrea' will be classified as woman when he is actually a man).

In their study, the country with the largest women's contribution to patenting is Spain (10.9%), followed by France (8.8%). Germany (3.5%) had the lowest percentage contribution.[2] Frietsch et al. (2009), using the same genderized names database as Naldi et al. (2004a, 2004b), extended the analysis to all EPO patents granted through 2005 for 14 countries. Again, while country-specific issues are addressed, the potential migratory background of inventors are not accounted for. Results point to Spain and France as the top countries in terms of gender balance (12.3% and 10.2% women's contribution in 2003-2005, respectively), and Germany and Austria closing the ranking. The US ranked toward the top, with 8.3% women contribution to patenting in 2003-2005. By technology fields, Pharmaceuticals (21%) and Basic chemicals (11.1%) stand out as the sectors with the largest women's contribution in 2005, with Energy machinery (2.2%) and Machine-tools (1.8%) ranked at the bottom.

Note that in all of these studies the inventors were previously not disambiguated. This means that women's shares could be downward biased if female inventors are positively selected – only the more capable women pursue an inventor career – and they are individually more productive than their male

[1] Alternative approaches to account for women patenting include, e.g., matching inventors' records with Social Security registers (Jung and Ejermo, 2014), surveys of inventors (Hoisl and Mariani, 2016; Walsh and Nagaoka, 2009), or semantics of how individuals are named (e.g., honorific titles or names structures: men's names ending with "o" and women's names ending with "a").

[2] Contribution to patents is measured by fractionalized counting of inventorships by gender. That is, each inventor is given 1/x count of the inventorships where x represents the number of inventors for which a gender could be assigned on the given patent.

counterparts. Or they could be upward biased if, for instance, female inventors have less experience than male inventors and therefore appear in less patent documents.

Another large-scale study by Sugimoto et al. (2015), using the entire universe of USPTO patents (1976-2013), determined inventor gender by comparing inventor names with several worldwide name information lists (e.g. Wikipedia and WikiName), country-specific gender information lists (e.g., US census), plus additional semantics for a few controversial countries such as Japan, Republic of Korea, India, China, etc. Their methods assigned gender to 90.8% of inventor names in the US, with attribution rates greater than 90% in other countries (Germany, the United Kingdom, Italy and Australia). Attribution rates were lower for inventors residing in India (66.1%), Republic of Korea (66.4%), Taiwan (72.9%), and the Netherlands (72.9%). According to their estimates, women accounted for 8% of patents for the whole period and 10.8% for the most recent year (2013).[3] Female patenting tended to be higher in university assigned patents and in chemistry fields. Neither disambiguation of inventors nor their migratory background is considered in Sugimoto's et al. (2015) study.

In a recent worldwide analysis released by the UK Intellectual Property Office (UKIPO), the authors of the study (UKIPO, 2016) use the entire universe of patents extracted from the EPO Worldwide Patent Statistics database, PATSTAT, and provide, to date, the largest existent inventor gender analysis. The authors use two external sources of genderized names: the first was the annual birth data from the US Social Security Administration and the UK Office for National Statistics (ONS). Second, they extracted name-gender list from millions of Facebook public profile pages (Tang et al., 2011). While the authors have undertaken a daunting effort to genderize listed inventors on all patents worldwide, 1915 to 2015, there are still some remaining issues. These issues include inventor disambiguation and, more importantly, the use of country of residence to infer inventor origin, ignoring potential biases caused by high-skilled migration flows. Additionally, inventors' residence is absent for about half of patents in PATSTAT. Attribution rates were roughly 80-90% for the US, Japan, the UK, Germany, France and Italy, but rates declined to around 75% for Switzerland and the Netherlands. Rates were even lower for China (27.90%), Republic of Korea (29.09%), and Taiwan (11.62%). Their results indicated that France, China and Russia (including the USSR) have the highest proportion of female inventors among the top inventor countries, with Japan, Germany and Republic of Korea having the lowest (the worldwide average is 7.2%, and the US share is 8.72%). Their analysis over time also revealed that women's contribution has significantly increased in recent years (2-4% during most of the 20th Century, 6% around the 2000s, and more than 11% in 2015). Additionally, the share of female inventors by fields again shows a large proportion for women patenting in Chemistry fields (between 10-20%) and Instruments: Analysis of biological materials (18.2%).

In another recent, worldwide study by Martinez et al. (2016), the authors built a dictionary of names for 182 different countries to classify the gender of inventors listed on Patent Cooperation Treaty (PCT) patents. Their list of 6.2 million names was built using country specific information sources, including Social Security registers and National Statistical Offices of various countries, Wikipedia lists, and incorporating manual checks by officials from the World Intellectual Property Organization (WIPO). Using this method, attribution rates were considerably higher than previous studies. Most rates were over 90% (97% for the US, 94% Japan, 99.2% Germany, 92.1% Republic of Korea, 98.9% the UK), and 88.3% in China and 88.9% in India. Their analysis revealed that women's contribution to patenting, measured again by fractional counting, had increased over time worldwide, from 9.5% in 1995 to around 15% in 2015 (the largest among all studies reviewed here). For the period of 2011-2015, the women contribution to patenting was high at 27-29% in the Republic of Korea and China, followed by 23% in Singapore, Spain and Poland and 13.9% in

[3] Again, measured by fractionalized counting of inventorships by gender.

the US. On the opposite side of the spectrum, the rate was as low as 7-9% in Austria and Germany. Women's patenting contribution was significantly larger in Chemistry fields, as well as in the field of Analysis of biological materials (Instruments). Differences were also apparent between academic and business patents (the former showing the largest women's shares in all countries analyzed). Interestingly, while their analysis does not provide genderized figures at the level of inventors (inventors are not disambiguated), the study is one of the few attempting to account for the potential migratory background of inventors by using nationality, and not country of residence, as the country of the inventors. This is possible thanks to the existence of nationality information of PCT inventors, at least until 2013 (Miguelez and Fink, 2013). Even if imperfect (some inventors may have acquired the nationality of their country of residence after some years), this is an important step in an era where migration flows, especially among high skilled workers, are on the rise – recent studies estimate the share of foreign inventors in the US to be around 20-24% (Kerr et al., 2016; Miguelez and Fink, 2013).

In recent research, Murray, Mariani and Delgado use data on the U.S. Social Security Administration and MIT students and applicants combined with USPTO data of inventor names to associate first names with gender, and in doing so create a state-by-state measure of Female Inventor Inclusivity. The authors find that the percent of female inventors in the US is 8.2%, with the highest percentages of female versus male inventors found in the Northeast and Midatlantic regions (New York City has 10.9% female inventors). The authors also provide a breakdown of female inventors patenting into their relevant technology classes, with the Drugs and Medical technology class patents leading the way with 30.3% female patents with at least one female inventor and the chemical technology class with 24.9%.

A.1.2. Gender attribution methods

In order to attribute gender to USPTO inventors, we use two sources of information:

1) The Global Name Recognition (IBM-GNR), a name search technology produced by IBM. IBM-GNR is a commercial product that performs various name disambiguation tasks. Among such tasks, we use two features. First, the association of names and surnames to one or (more often) several countries of likely origin. Second, the association of names to male and female gender and their associated probabilities. These associations originate from a database produced by the US immigration authorities in the first half of the 1990s, which registered all names and surnames, along with nationality and gender, of all foreign citizens entering the US. It contains a total of roughly 750,000 full names. In addition, variants of registered names and surnames are considered, according to country-sensitive orthographic and abbreviation rules – more information in Breschi et al. (2017a, 2017b).

2) The WIPO worldwide gender-name dictionary (WGND), produced by the World Intellectual Property Organization (WIPO). It includes a list of 6.2 million names from 182 different countries. For each name contained in the dataset, it attaches a given gender to each name by country where that name appears in the original source data. The construction of the WGND draws on previous gender studies (see section A.1.1) as well as from national public statistical institutions (plus an ad-hoc list of names created by WIPO staff) – see Martínez et al. (2016) for details. For several names in certain countries, the name is used for both female and male, and therefore attribution for that given name in a given country is left as 'unknown'.

Using these two sources of information, we attribute gender to USPTO inventor names following these steps:

1) For each inventor name, IBM-GNR returns the share of instances it identifies as male in the data source and the share it identifies as female. It also returns an additional metric ("frequency"), which indicates the frequency percentile that each name belongs within the complete dataset. A very uncommon name will be assigned a very low frequency (possibly even zero if it doesn't exist in the dataset). Therefore, gender attribution for low frequency values is assumed to be less reliable.

2) For each inventor first name, we attribute female gender to a given inventor if it is identified as female in 97% or more cases and we attribute male gender to a given inventor if it is identified as male in 98% or more cases. Threshold values were decided by visual inspection of the distribution of shares. However, we do not attribute any gender to those names which are very rare in IBM-GNR dataset, that is, those names with a frequency of 5% or less. This step resulted in gender attribution to 71.79% of inventors (2,499,999 inventors).

3) When the inventor name is majority female (but not in the 97% of the cases) and the second (or middle) name is 97% or more female, we also attribute female gender. Similarly, when the inventor name is majority male (but not in the 98% of the cases) and the second name is 98% or more male, we attribute male gender. Doing this we attribute 38,581 additional names, reaching 72.90% of attribution.

4) For the remaining 943,725 inventor names, we rely on WIPO's WGND. As mentioned previously, WGND is a dictionary of 6.2 million names associated with 182 different countries. For each name+country pair, we can match the inventor name to the WGND to determine whether the name has predominantly female or male attribution within that country. To do this, we first need to assign a country of origin to each inventor. Historically, the language spoken by the inventor (or, better, by his/her parents, who chose that name) could be safely inferred from the individual's country or region of residence. Unfortunately, as discussed in section 1, residency is decreasingly signifies of origin, due to the extent of and continuous growth in highly skilled labor migration, both to the US and other advanced countries. We thus come back to GNR to determine likely country of origin of the inventors.

5) As previously discussed, GNR associates inventor names and surnames to a vector of likely country/ies of origin. The US is never listed among these countries of origin because the raw data is extracted from the US immigration authorities. However, other Anglo-Saxon countries (the UK, Ireland, Australia, Canada, etc.) are included. For each likely country of origin, GNR attaches a "significance" measure, which indicates the share of instances the name or surname is associated with a given country of origin. For the purpose of the present algorithm, we focus only on the vector of countries associated to inventor surname in order to determine the likely country of origin. This is done because we are interested in determining the gender of the first name and therefore it cannot be part of our decision rule. Imagine the case, again, of "Andrea". This name is likely to be Italian, but it is even more likely to be from Spanish-speaking countries. If the surname was mostly Italian and we used the inventor first name, which has large percent share from Spanish-speaking countries, we could run the risk of assigning to this inventor a Spanish-speaking origin, while he is clearly Italian. Unfortunately, this decision also creates problems in countries where women adopt the surname of their husband upon marriage. However, we posit that most marriages tend to be intra-ethnic, so would not affect many cases.

6) Of the vector of countries associated to the surname of the inventor, we focus only on those with at least a 10% significance. We then sort the countries of origin, from the country with largest significance to smallest (over 10%). Note an additional problem, best understood with an example:

"Smith". This surname could be associated to Germany with a significance measure of 30%, the UK with 20%, Ireland with 10%, and Australia with 10%. In principle, we would associate "Smith" primarily to Germany. However, if we aggregate all the Anglo-Saxon shares, they add up to 40%, which is larger than the German share. To address this, some countries are collapsed into linguistic groups to create the vector of countries/languages associated to surnames and then sorted by their share. (Table A.1.3 presents a list of the country linguistic groups).

7) Each inventor is then associated with a vector of linguistic groups or individual countries, in descendent order according to the significance measure. Thus, following on with the example above, each inventor whose surname is "Smith" will first be associated with the UK, Ireland, the US, Australia, Canada, Bermuda, etc. (all English-speaking countries), and then to Germany, Switzerland and Austria.

8) With the linguistic group or country associated to each inventor (and the order in which they are associated), we match the first name of the inventor and at least one of the associated countries to the name+country pairs in the WGND dataset. We keep more than one linguistic group per inventor because, for some names-countries, the first linguistic group does not exist in the WGND dataset. In those cases, we use the second linguistic group (or the third if the second is also absent, and so on).

9) For some inventors, with rare surnames, we were not able to create a vector of likely countries of origin. For these cases, we used the country of residence – 37,003 total number of cases which was 3.92% of the 943,725 inventor names.

10) In the end, the match with WGND resulted in an additional 498,620 inventor name gender attributes.

11) Lastly, for some cases (169,405 total inventor names, 18% of the 943,725 inventor names), no name+country match exists in the WGND. For these cases, we still attribute gender if (1) the first name appears in the WGND dataset (although not linked to the country/ies of the inventor) and if it is male or female in all cases; and (2) the male or female attribution from WGND coincides with the majority of instances attributed by GNR.

After conducting these 11 steps ("baseline" method), we were able to attribute gender to 3,206,605 inventors, 92.08% of all USPTO inventors (7.92% of non-attributed cases; see table A.1.1).

As shown in tables A.1.1 and A.1.2, even though the US is quantitatively the country with more non-attributed inventor names, the non-attribution rate is much higher in countries such as China, India, and the Republic of Korea. We share this problem with similar, prior studies that have attempted to assign gender to Asian names. We implement, therefore, some additional steps ("baseline-augmented" method):

12) For inventors whose surname is primarily associated with China, Singapore, Taiwan, Macao, or Hong Kong (even when they do not reside in those countries), we attribute female gender if it is identified as female in 60% or more cases, and we attribute male gender if it is identified as male in 60% or more cases (threshold decided upon visual inspection of the distribution of GNR's shares).

13) For inventors whose surname is primarily associated to the Republic of Korea (even when they do not reside in that country), we attribute female gender if it is identified as female in 80% or more cases, and we attribute male gender if it is identified as male in 80% or more cases (threshold decided upon visual inspection of the distribution of GNR's shares).

14) For inventors whose surname is primarily associated to India (even when they do not reside in that country), we attribute female gender if it is identified as female in 90% or more cases, and we attribute male gender if it is identified as male in 90% or more of the cases (threshold decided upon visual inspection of the distribution of GNR's shares).

After conducting these remaining steps, we attributed gender to 38,188 additional inventors. In total, our (baseline-augmented) method attributes gender to 3,244,813 inventors, 93.18% of all USPTO inventors (6.82% of non-attributed cases) – broken down by countries in column 2 of tables A.1.1 and A.1.2.

Tables A.1.1 and A.1.2 also show non-attribution shares under two alternative approaches, namely:

- The "ethnic-based" approach, which uses only the IBM-GNR worldwide list to attribute gender (e.g., a given name is feminine if IBM-GNR attributes it to be feminine in 51% of the cases);

- The "ethnic-based-augmented" approach, which uses the IBM-GNR method and the WGND's country specific list for the cases that do not appear in IBM-GNR.

As shown below, the "ethnic-based-augmented" approach performs better than our preferred "baseline-augmented" approach. However, precision is likely affected which means that the gender attribution is less reliable.

Table A.1.1: Gender non-attribution cases (% on total inventors worldwide and country breakdown)

	Baseline		Baseline-augmented		Ethnic-based		Ethnic-based-augmented	
	%	**# (,000)**	**%**	**# (,000)**	**%**	**# (,000)**	**%**	**# (,000)**
All countries	7.92	276	6.82	237	9.44	329	5.45	190
	of which:		of which:		of which:		of which:	
AT	0.01	0.3	0.01	0.3	0.01	0.5	0.01	0.2
AU	0.03	0.9	0.02	0.8	0.04	1.4	0.02	0.7
BE	0.02	0.6	0.02	0.6	0.02	0.8	0.01	0.4
BR	0.01	0.5	0.01	0.5	0.02	0.6	0.01	0.4
CA	0.14	5	0.13	4.5	0.23	8.1	0.11	3.9
CH	0.03	0.9	0.03	0.9	0.05	1.6	0.02	0.7
CN	0.99	34.6	0.85	29.6	0.84	29.3	0.78	27.3
DE	0.12	4.3	0.12	4.2	0.26	9	0.11	3.7
DK	0.01	0.4	0.01	0.4	0.03	0.9	0.01	0.4
ES	0.01	0.3	0.01	0.3	0.02	0.6	0.01	0.3
FI	0.03	0.9	0.03	0.9	0.04	1.4	0.01	0.4
FR	0.10	3.6	0.10	3.6	0.10	3.6	0.06	2
GB	0.06	2.2	0.06	2.1	0.11	3.8	0.05	1.8
GR	0.00	0.1	0.00	0.1	0.01	0.2	0.00	0.1
HK	0.03	1.2	0.02	0.8	0.02	0.6	0.01	0.5
HU	0.01	0.3	0.01	0.3	0.02	0.7	0.01	0.3
IE	0.00	0.1	0.00	0.1	0.01	0.3	0.00	0.1
IL	0.05	1.7	0.05	1.7	0.10	3.4	0.04	1.5
IN	0.26	8.9	0.25	8.7	0.42	14.5	0.25	8.7
IT	0.04	1.4	0.04	1.4	0.07	2.6	0.04	1.3
JP	1.56	54.4	1.56	54.2	2.03	70.9	1.48	51.7
KR	0.81	28.3	0.66	23	0.32	11.3	0.20	6.9
MX	0.00	0.1	0.00	0.1	0.00	0.1	0.00	0.1
MY	0.04	1.3	0.03	1	0.02	0.7	0.01	0.5
NL	0.06	2	0.05	1.9	0.11	3.9	0.05	1.9
NO	0.01	0.5	0.01	0.4	0.02	0.7	0.01	0.4
NZ	0.00	0.1	0.00	0.1	0.01	0.2	0.00	0.1
PL	0.00	0.1	0.00	0.1	0.01	0.2	0.00	0.1
PT	0.00	0	0.00	0	0.00	0	0.00	0
RU	0.02	0.6	0.02	0.6	0.03	1.2	0.02	0.6
SE	0.03	1.2	0.03	1.2	0.05	1.7	0.02	0.8
SG	0.09	3.2	0.07	2.4	0.05	1.9	0.04	1.5
TR	0.00	0.1	0.00	0.1	0.01	0.4	0.00	0.1
TW	0.84	29.3	0.35	12.3	0.14	4.8	0.12	4.1
US	2.36	82.2	2.14	74.2	3.95	137.8	1.78	62
ZA	0.01	0.3	0.01	0.3	0.01	0.4	0.01	0.2

Table A.1.2: Gender non-attribution cases (% of total inventors by country)

	Baseline	Baseline-augmented	Ethnic-based	Ethnic-based-augmented
	%	%	%	%
AT	1.66	1.63	3.05	1.45
AU	3.23	2.94	5.2	2.52
BE	3.24	3.17	4.56	2.41
BR	6.84	6.84	8.7	5.99
CA	5.21	4.74	8.55	4.05
CH	2.5	2.44	4.42	2
CN	62.34	53.33	52.8	49.22
DE	1.6	1.57	3.36	1.39
DK	3.17	3.12	6.09	2.92
ES	1.95	1.92	3.58	1.66
FI	5.11	5.08	7.55	2.12
FR	3.05	3.01	3.05	1.65
GB	2.05	1.93	3.55	1.68
GR	6.92	6.92	19.47	6.75
HK	17.78	11.82	9.23	6.85
HU	4.38	4.38	9.9	4.08
IE	2.71	2.68	5.07	2.09
IL	5.81	5.79	11.79	5.28
IN	29.19	28.5	47.36	28.27
IT	3.16	3.15	5.76	2.95
JP	10.73	10.69	13.98	10.19
KR	30.75	24.99	12.3	7.5
MX	1.72	1.68	2.97	1.49
MY	36.88	28.63	19.56	14.79
NL	4.78	4.72	9.4	4.56
NO	5.11	5.07	8.01	4.25
NZ	3.07	2.92	4.55	2.32
PL	2.98	2.92	6.41	2.4
PT	1.88	1.88	2.04	1.31
RU	6.45	6.44	11.81	6.47
SE	2.99	2.91	4.29	2.06
SG	33.78	25.07	20.19	15.9
TR	7.01	7.01	35.38	6.85
TW	31.67	13.3	5.25	4.43
US	4.87	4.39	8.16	3.67
ZA	5.41	5.41	8.65	5.06
All countries	7.92	6.82	9.44	5.45

Table A.1.3: Linguistic groups

1	2	3	4	5	6	7	8	9	10
Country code	Linguistuic group	Country code	Linguistuic group	Country code	Linguistuic group	Country code	Linguistuic group	Country code	Linguistuic group
AD	Spanish	CW	CW	IQ	Arabic	MX	Spanish	SR	Ducth
AE	Arabic	CY	Greek	IR	Persian	MY	MY	SS	SS
AF	Persian	CZ	Slavic	IS	Occidental Scandinavian	MZ	MZ	ST	ST
AG	English	DE	German	IT	Italian	NA	NA	SU	Russian
AI	English	DJ	DJ	JE	English	NC	French	SV SX	Spanish
AL	AL	DK	Oriental Scandinavian	JM	English	NE	NE	SY	English
AM	AM	DM	English	JO	Arabic	NF	English	SZ	Arabic
AN	AN	DO	Spanish	JP	JP	NG	NG	TC	SZ
AO	AO	DZ EC	Arabic	KE	KE	NI	Spanish	TD TF	English TD
AR	Spanish	EE	Spanish	KG	Turkic	NL	Dutch	TG	TF
AS	AS	EG	Finnic	KH	KH	NO	Occidental Scandinavian	TH	
AT	German	EH	Arabic	KI	KI	NP	NP	TJ TL	TG
AU	English	ER	Arabic	KM	KM	NR	NR	TM	TH
AW	AW	ES	ER	KN	English	NU	NU	TN TO	Persian
AZ	Turkic	ET	Spanish	KP	Korean	NZ	English	TR	TL
BA	Serbo-Croatian	FI	ET	KR	Korean	OM	OM	TT	Turkic
BB	English	FJ	Finnic	KW	Arabic	PA	Spanish	TV	Arabic
BD	BD	FK	English	KY	English	PE	Spanish	TW TZ	TO
BE BF	Dutch	FM	English	KZ	Turkic	PF	French PG	UA	Turkic
BG	BF	FO	English	LA	LA	PG	PH	UG	English
BH	Orienatal	FR	Occidental Scandinavian	LB	Arabic	PH	PK	US	TV
BI	Slavic	GA	French	LC	English	PK	Slavic	UY	Chinese
BJ	Arabic BI	GB	GA	LI	German	PL	French	UZ	TZ
BM	BJ	GD	English	LK	LK	PM	English	VA	Russian
BN	English	GE	English	LR	LR	PN	Spanish	VC	UG
BO	Malay	GF	GE	LS	LS	PR	Arabic	VE	English
BQ	Spanish	GG	French	LT	Baltic French	PS	Portuguese	VG	Spanish
BR	BQ	GH	English	LU	Baltic	PT	PW	VI	Turkic
BS BT	Portuguese	GI	GH	LV	Arabic	PW	Spanish	VN	Italian
BW	English	GL	Spanish	LY	Arabic	PY	Arabic	VU WF	English
BY BZ	BT	GM	GL	MA	French	QA	French	WS	Spanish
CA	BW	GN	GM	MC	MD	RE	RO	YE	English
CD	Russian	GP	GN	MD	Serbo-Croatian	RO	Serbo-Croatian	YU	English
CF	English	GQ	GP	ME	French	RS	Russian	ZA	VN
CG	English CD CF	GR	GQ	MF	MG	RU	RW	ZM	VU
CH	CG	GT	Greek	MG	MH	RW	Arabic	ZW	WF
CI	German	GU	Spanish	MH	Oriental	SA	SB		WS
CK	CI	GW	GU	MK	Slavic	SB	SC		Arabic
CL	English	GY	GW	ML	ML	SC	SD		Serbo-Croatian
CM	Spanish	HK	English	MM	MM	SD	Oriental Scandinavian		ZA
CN	CM	HN	Chinese	MN	MN	SE	Chinese		ZM
CO	Chinese	HR	Spanish	MO	Chinese	SG	SH		ZW
CR	Spanish	HT	Serbo-Croatian	MP	MP	SH	Serbo-Croatian		
CS	Spanish		French	MQ	MQ	SI	Occidental Scandinavian		
CU	Slavic	HU	HU	MR	Arabic	SJ	Slavic		
CV	Spanish	ID	Malay	MS	MS	SK	SL		
CX	CV	IE	English	MT	MT	SL	Italian		
	CX	IL	IL	MU	MU	SM	SN		
		IM	English	MV	MV	SN	SO		
		IN	IN	MW	MW	SO			

Note: Country codes are presented in odd columns, and their corresponding linguistic group in even columns. If even columns contain a country code, instead of a linguistic group, it means that no action was taken and that country was not grouped with other linguistically similar ones. In order to identify relevant linguistic groups, we used several sources, including CIA Factbook 2016, Ethnologue (https://www.ethnologue.com/), Wikipedia, and CEPII (*Centre d'Etudes Prospectives et d'Informations Internationales*, in particular, see: Melitz and Toubal, 2014)

Appendix 2: Additional descriptive tables

Figure A.2.1: Women inventor rates, by country

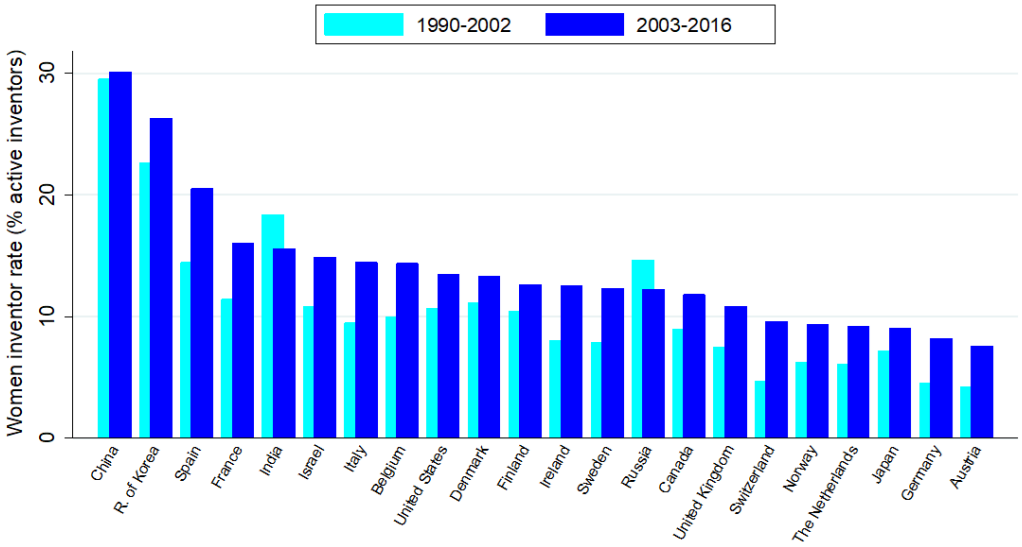

Figure A.2.2: Women inventor rate across US CBSAs, 2003-2016

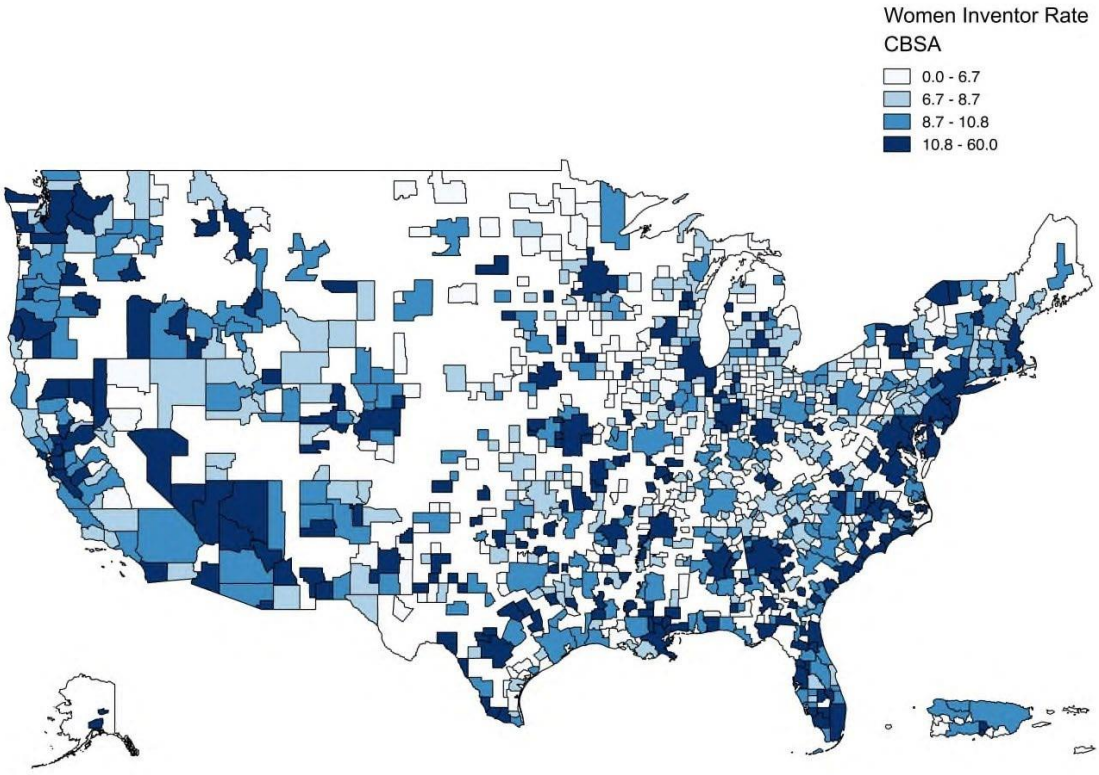

Figure A.2.3: Women inventor rate across fields, 2003-2016

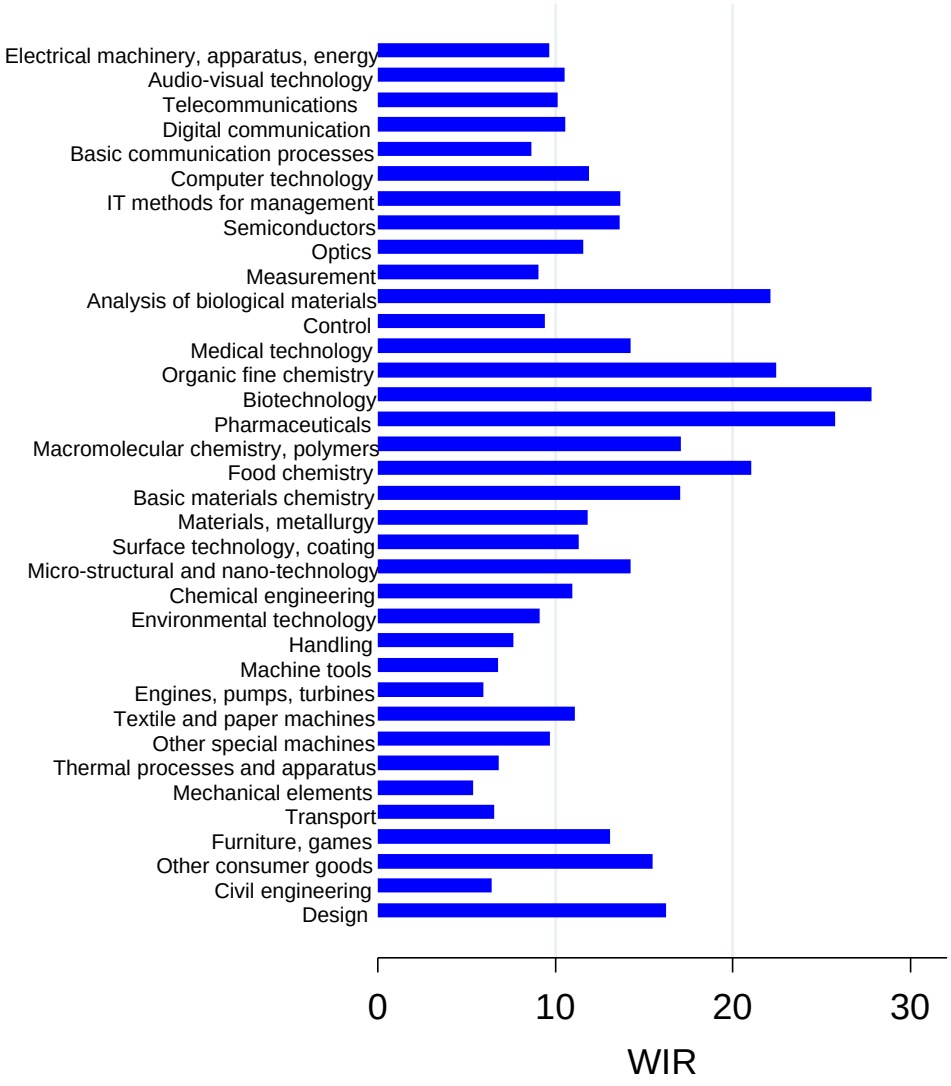

Table A.2.1: WIPO technology classification of patents' IPC codes

Sector	Field
Electrical engineering	Electrical machinery, apparatus, energy
Electrical engineering	Audio-visual technology
Electrical engineering	Telecommunications
Electrical engineering	Digital communication
Electrical engineering	Basic communication processes
Electrical engineering	Computer technology
Electrical engineering	IT methods for management
Electrical engineering	Semiconductors
Instruments	Optics
Instruments	Measurement
Instruments	Analysis of biological materials
Instruments	Control
Instruments	Medical technology
Chemistry	Organic fine chemistry
Chemistry	Biotechnology
Chemistry	Pharmaceuticals
Chemistry	Macromolecular chemistry, polymers
Chemistry	Food chemistry
Chemistry	Basic materials chemistry
Chemistry	Materials, metallurgy
Chemistry	Surface technology, coating
Chemistry	Micro-structural and nano-technology
Chemistry	Chemical engineering
Chemistry	Environmental technology
Mechanical engineering	Handling
Mechanical engineering	Machine tools
Mechanical engineering	Engines, pumps, turbines
Mechanical engineering	Textile and paper machines
Mechanical engineering	Other special machines
Mechanical engineering	Thermal processes and apparatus
Mechanical engineering	Mechanical elements
Mechanical engineering	Transport
Other fields	Furniture, games
Other fields	Other consumer goods
Other fields	Civil engineering
Design	Design

Appendix 3: Cross-country variation and the composition effect

Table A.3.1. reports the results of a set of binary logit regressions, by means of which we estimate the probability for an inventor chosen at random from PatentsView to be a woman, controlling for a set of country dummies along with dummies for sectors and types of applicants (and their interactions) – plus time fixed effects. The unit of observation is the inventor with at least one patent in the years from 2003 to 2016. We further restrict the sample to inventors with the residence, at the time of their first patent, in the US, Japan, China, the Republic of Korea and the three largest countries of the EPO (Germany, France, and the UK) – IP5, which implies a total of 1,435,741 inventors. The technology to which we assign inventors correspond to the technological classes of her patents (five macro classes, according to Schmoch's (2008) and WIPO's classification of IPC codes, plus design and plant variety patents). When the inventor has patents in more than one sector, all them sectors are considered, so that technology dummies are not mutually exclusive. In the same vein, the type of assignee refers to the owner of the patent. However, contrary to the sector dummies, assignee dummies are mutually exclusive: the dummy *UNIVERSITY* is valued 1 if the inventor has at least 1 patent assigned to a university or hospital. The dummy *PRO* is valued 1 if she has at least a patent applied by a Public Research Organization (PRO) and no one by universities. The dummy *BUSINESS* is valued 1 if she has at least a patent applied by a firm and no patent applied neither by universities nor by PROs. Finally, the dummy INDIVIDUAL is valued 1 in the remaining cases.[4] The year dummies refer to the year of the first patent produced by the inventor in the years from 2003 to 2016.

In all regressions, the estimated parameters are the logit coefficients, which may be also interpreted as odds ratios (ORs=$e^{coeff.}$). Odds are defined as the ratio of the probability of success (the inventor being woman) and the probability of failure (the inventor being man). Odds range between 0 and infinity: odds ratio greater than one describes a positive relationship. The US is the reference case, so that the odds ratios for country dummies represent marginal variations for the odds of the inventor being a woman in any specific country, relative to the US.

In column 1 we estimate the probability that an inventor is woman only as a function of country dummies. For China, the Republic of Korea and France, the estimated coefficients are positive, meaning that in these countries the probability of being woman is higher than in the US. In particular, in terms of odds ratios, the relative probability for a French inventor to be a woman instead of a man is 1.25 (=$e^{0.220}$) times greater than the analogous probability for a US inventor. Without any further controls on, i.e., technological fields or type of assignees, this number reflects the observed shares of women in US and in France: in the US the share of women is approximately 13%, corresponding to an odds ratio of 0.15 (13%/87%), while in France the share is approximately 16%, corresponding to an odds ratio of 0.19 (16%/84%). As for the other selected countries, China shows the largest odds ratio. In particular, the odds ratio of China is 2.81; meaning that being a Chinese inventor puts you at 2.81 (=$e^{1.032}$) times greater odds of being woman with

[4] While classifying patents by technology is a straightforward operation, the opposite is not true for types of assignee, for which patent authorities do not maintain nor apply any reliable classification system. Hence, we adopted the classification proposed by Van Looy (2006) and Du Plessis et al. (2009), who produce and maintain the EEE-PPAT dataset, nowadays directly available through PATSTAT. The method is based on string analysis of patent assignees' names, where it searches for keywords such as "University", "Research Institute", "Government", "Hospital", "Limited", "Inc." and so forth. We connected the patents in PatentsView and in PATSTAT by means of their publication numbers. For patents co-assigned to different types of organizations we proceeded by classifying to Universities & Hospitals, if at least one these organizations appeared among the co-assignees, moving then to consider, sequentially, the presence among co-assignees of a PRO or a Business companies. Patents classified as assigned to Individuals have no organization of sort among the co-assignees.

respect to a US inventor. Japan, Germany and the UK show negative coefficients, meaning that in these countries the probability to observe a woman inventor is lower than in the US.

In columns 2 to 4 we introduce several controls in a stepwise fashion, which capture the composition effect discussed in section 3.5 of the main report. To the extent that such composition effects explain the observed cross-country differences with respect to WIR, we expect the introduction of such controls to reduce the estimated Odds Ratios for the country dummies.

In column 2 we control for the inventors' age, which we approximate with the year dummies corresponding to the inventors' first patenting year. In column 3 we add technology fixed effects. Finally, in column 4 we also add the assignee type dummies and some interactions between fixed effects, which should control for all kinds of time-varying and sector-varying unobservable characteristics. As expected, we find that differences between the US and the countries showing higher observed WIR values decrease notably. The odds ratio for China declines to 1.96 ($=e^{0.672}$) from 2.81, and those from 2.32 ($=e^{0.672}$) to 1.68 for the Republic of Korea. The odds radio associated for France becomes 1.02 and not significant, meaning that, controlling for composition effects, a French inventor has the same probability of a US inventor of being a woman. In other terms, the higher observed prevalence of women in France with respect to the US is likely due to the higher share of patents in universities, hospitals and PROs (see Figure 10). With respect to the countries with negative coefficients, the differences persist, although they are also significantly reduced. The estimation results for column 4 are summarized in figure 12 of the main report.

Table A.2.1 - Probability of WIR and composition effects: US, France, Germany, UK, Japan, Republic of Korea and China (2003-2016)
(Logit regressions, - Coefficients, and standard errors in brackets)

	Logit	Logit	Logit	Logit
	(1)	(2)	(3)	(4)
Country				
(Reference: United States)				
China	1.032***	0.910***	0.896***	0.672***
	(0.0155)	(0.0157)	(0.0163)	(0.0426)
Germany	-0.548***	-0.560***	-0.551***	-0.420***
	(0.0102)	(0.0102)	(0.0102)	(0.0356)
France	0.220***	0.187***	0.132***	0.0244
	(0.0116)	(0.0117)	(0.0117)	(0.0437)
Japan	-0.439***	-0.413***	-0.371***	-0.255***
	(0.00749)	(0.00751)	(0.00767)	(0.0198)
South Korea	0.842***	0.788***	0.818***	0.517***
	(0.0103)	(0.0104)	(0.0108)	(0.0266)
UK	-0.240***	-0.254***	-0.350***	-0.165**
	(0.0145)	(0.0145)	(0.0145)	(0.0540)
Type of Assignee				
(Reference: BUSINESS)				
INDIVIDUAL				0.597***
				(0.0505)
PROs				0.201***
				(0.0553)
UNIVERSITIES				0.227***
				(0.0393)
Constant	-1.880***	-2.207***	-1.827***	-1.921***
	(0.00323)	(0.00768)	(0.0112)	(0.0144)
Year FE	NO	YES	YES	YES
Technology FE	NO	NO	YES	YES
Technology x Country FE	NO	NO	NO	YES
Technology x Assignee FE	NO	NO	NO	YES
Year x Assignee FE	NO	NO	NO	YES
Observations	1,435,741	1,435,741	1,435,741	1,426,636

Notes: Robust standard errors in parentheses: *** p<0.001, ** p<0.01, * p<0.05. Year FEs refer to the first patenting year of the inventor patenting history (in the years 2003-2016). Inventor are assigned to a country according to the address reported on their first patent in time

References

Breschi, S., Lissoni, F., Miguelez, E., 2017a. Foreign-origin inventors in the USA: testing for diaspora and brain gain effects. J Econ Geogr 17, 1009–1038. https://doi.org/10.1093/jeg/lbw044

Breschi, S., Lissoni, F., Tarasconi, G., 2017b. Inventor Data for Research on Migration & Innovation: The Ethnic-Inv Pilot Database., in: In: FINK, C. & MIGUELEZ, E. (Eds.) The International Mobility of Talent and Innovation: New Evidence and Policy Implications. Cambridge University Press.

Delgado, M., Mariani, M., Murray, F., 2018. The Role of Location on the Inventor Gender Gap. GEOINNO2018, 4th Geography of Innovation Conference. MIT Management Sloan School and MIT innovation initiative.

Delgado, M and F. Murray, 2019, "Catalysts for Gender Inclusion in Innovation: The Role of Universities and their Top Inventors," Working Paper, MIT

Frietsch, R., Haller, I., Funken-Vrohlings, M., Grupp, H., 2009. Gender-specific patterns in patenting and publishing. Research Policy, Special Issue: Emerging Challenges for Science, Technology and Innovation Policy Research: A Reflexive Overview 38, 590–599. https://doi.org/10.1016/j.respol.2009.01.019

Hoisl, K., Mariani, M., 2016. It's a Man's Job: Income and the Gender Gap in Industrial Research. Management Science 63, 766–790. https://doi.org/10.1287/mnsc.2015.2357

Jung, T., Ejermo, O., 2014. Demographic patterns and trends in patenting: Gender, age, and education of inventors. Technological Forecasting and Social Change 86, 110–124. https://doi.org/10.1016/j.techfore.2013.08.023

Kerr, S.P., Kerr, W., Özden, Ç., Parsons, C., 2016. Global Talent Flows. Journal of Economic Perspectives 30, 83–106. https://doi.org/10.1257/jep.30.4.83

Martinez, C., Ciaramella, L., Ménière, Y., 2016. How to track patent transfers in Europe: a first empirical analysis. http://www.epip2016.org/book-of-abstracts/martinez.

Martínez, G.L., Raffo, J., Saito, K., 2016. Identifying the Gender of PCT inventors (No. 33), WIPO Economic Research Working Papers. World Intellectual Property Organization - Economics and Statistics Division.

Melitz, J., Toubal, F., 2014. Native language, spoken language, translation and trade. Journal of International Economics 93, 351–363. https://doi.org/10.1016/j.jinteco.2014.04.004

Miguelez, E., Fink, C., 2013. Measuring the International Mobility of Inventors: A New Database (WIPO Economic Research Working Paper No. 8).

Naldi, F., Luzi, D., Valente, A., Parenti, I.V., 2004a. Scientific and Technological Performance by Gender. SpringerLink 299–314. https://doi.org/10.1007/1-4020-2755-9_14

Naldi, F., Luzi, D., Valente, A., Parenti, I.V., 2004b. Scientific and technological performance by gender. Handbook of quantitative science and technology research 299–314.

Schmoch, U., 2008. Concept of a technology classification for country comparisons. Final report to the World Intellectual Property Organization (WIPO), Fraunhofer Institute for Systems and Innovation Research, Karlsruhe.

Sugimoto, C.R., Ni, C., West, J.D., Larivière, V., 2015. The Academic Advantage: Gender Disparities in Patenting. PLOS ONE 10, e0128000. https://doi.org/10.1371/journal.pone.0128000

Tang, C., Ross, K., Saxena, N., Chen, R., 2011. What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook, in: Database Systems for Adanced Applications, Lecture Notes in Computer Science. Presented at the International Conference on Database Systems for Advanced Applications, Springer, Berlin, Heidelberg, pp.344–356. https://doi.org/10.1007/978-3-642-20244-5_33

UKIPO, 2016. Gender profiles in worldwide patenting: An analysis of female inventorship.

Walsh, J.P., Nagaoka, S., 2009. Who Invents?: Evidence from the Japan-U.S. inventor survey.