Discrimination against foreigners in the U.S. patent system

Gaétan de Rassenfosse
Reza Hosseini

September 2020

Innovation and Intellectual Property Policy
Working Paper series no. 12

Available at: https://ideas.repec.org/p/iip/wpaper/12.html

Discrimination against foreigners in the U.S. patent system

Gaétan de Rassenfosse[*],✉ and Reza Hosseini[*]

[*] Chair of Innovation and IP Policy, College of Management of Technology,
Ecole polytechnique fédérale de Lausanne. Station 5, 1015 Lausanne, Switzerland.

✉ Corresponding author: gaetan.derassenfosse@epfl.ch

ABSTRACT

Inventions of foreign origin are about ten percentage points less likely to be granted a U.S. patent than domestic inventions. An empirical analysis of 1.5 million U.S. patent applications identifies three systematic differences between foreign and domestic patent applications that partly explain this bias. They include differences in patent agents, financial resources of the applicants, and the level of effort that applicants put into the prosecution process. We find no evidence of disparate treatment ('intentional discrimination') of foreigners. Instead, our evidence points to a disparate impact ('unintentional discrimination') of the U.S. patent system on foreign inventors. Our results suggest unequal access to the patent system for foreigners compared to locals (but also for small U.S. firms). Giving examiners the power of (truly) rejecting a patent application may be one solution to level the playing field between foreigners and locals, but also between large and small firms.

Keywords: foreign bias; discrimination; disparate impact, national treatment principle; patent system

> "Each Member shall accord to the nationals of other Members treatment no less favourable than that it accords to its own nationals with regard to the protection of intellectual property […]"
>
> TRIPS Agreement, Article 3.1

INTRODUCTION

The 'national treatment principle,' which imposes equal treatment of foreigners and locals, is a fundamental aspect of international patent law. It was established by the 1883 Paris Convention for the Protection of Industrial Property, and it has been reaffirmed recently with the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) by the members of the World Trade Organization.

The national treatment principle is the *raison d'être* of the global patent system and underpins international business. Innovative multinational enterprises (MNEs) would be reluctant to seek patent protection in foreign jurisdictions if local patent offices were allowed to discriminate against them. In turn, a weakening of patent protection would hurt international trade—indeed, scholars generally consider strong patent rights as beneficial to international trade (Maskus and Penubarti 1995, Ivus 2015, Palangkaraya *et al.* 2017). Furthermore, discrimination against foreign MNEs would lower the returns to inventive activities, putting at risk the ability of the global patent system to stimulate R&D investment.

There is mounting evidence of antiforeign bias in patent systems (Kotabe 1992, Popp *et al.* 2003, Yang 2008, Harhoff and Wagner 2008, Liegsalz and Wagner 2013, Webster *et al.* 2014, Tong *et al.* 2018, Yang and Sonmez 2018, Yang 2019, de Rassenfosse *et al.* 2019). However, most studies offer correlational, instead of causal, evidence, such that it is difficult to conclude that the national treatment principle is not being upheld. As far as we can ascertain, Webster *et al.* (2014) were the first to offer a counterfactual analysis. The authors analyzed a set of inventions that were all granted patent protection at the U.S. Patent and Trademark Office (USPTO) and submitted for protection at both the European Patent Office (EPO) and the Japanese Patent Office (JPO). They found that European inventors were more likely to have their patent applications granted in Europe than Japanese inventors *ceteris paribus* (and vice-versa for Japanese inventors at the JPO). Webster and colleagues took this result as evidence that the patent system treats foreigners unfavorably, in apparent violation of international patent law. de Rassenfosse *et al.* (2019) adopt a similar approach and offer more recent evidence of potential discrimination at the five largest offices. So far, however, the literature has remained silent on the channel(s) through which this bias against foreigners manifests itself.

The present paper contributes to this nascent literature by investigating possible channels for the antiforeign bias. The empirical analysis models the grant outcome of about 1.5 million U.S. patent applications filed in the years 2002–2012. We adopt recent methodological advances in the field by exploiting information on the grant outcome of a set of twin patents filed in foreign jurisdictions, which we use to control for the likelihood of grant of the U.S. applications.

We find that inventions of foreign origin are about ten percentage points less likely to be granted a patent than domestic inventions, which suggests discrimination against foreigners. This discrimination could be intentional or unintentional. Intentional discrimination relates to disparate treatment of a specific group of applicants, whereas unintentional discrimination arises when policies, practices, and rules have disparate impacts on a specific group of applicants. We show that bias against foreigners is largely the result of unintentional discrimination. It can be explained by differences in patent agents between foreigners and locals, the financial resources of the applicants, and the level of effort that applicants put into the prosecution process.

BACKGROUND

Understanding discrimination

Concerns about 'discrimination' against foreigners in the patent system are not new. A major point of tension in the 1990s on this issue concerned Japan (*e.g.*, Helfgott 1990), and more recent discussions have focused on China (*e.g.*, Harris 2009, Brander *et al.* 2017). As regards to empirical evidence, Kotabe

(1992:147) observes the aggregate grant rate at the leading patent offices and concludes that "U.S., German, and British patent practices appear to discriminate against foreign applicants with lower patent grant ratios than for domestic applicants." Yang (2008:1035) observes that the grant rate at China National Intellectual Property Administration (CNIPA) is higher for domestic than for foreign applicants and concludes that "China appears to give preferential treatment to domestic applications."

An essential semantic clarification is in order before proceeding. The finding that foreigners have lower grant rates than locals does not in itself suggest 'discrimination' or 'bias.' Indeed, there might be legitimate reasons for aggregate differences in grant rates between foreigners and locals. The ideal experiment for testing discrimination would be to assign randomly foreign and local origins to a set of patent applications submitted to a patent office. Since patent applications would be otherwise similar on average, systematic differences in the grant rates between patent applications by foreigners and locals would provide evidence of discrimination. Bertrand and Mullainathan (2004) have adopted a similar approach to test for discrimination in the labor market.

Unfortunately, the patent system does not lend himself well to such an experiment owing to the cost and complexity of the patenting process. Scholars have relied instead on observational data, which requires a detailed understanding of the determinants of grants in order to control for confounding factors. The single most important determinant of a grant is admittedly the 'quality' of the patent application. Patent quality is a multifaceted concept that scholars have discussed at length (*e.g.*, Guerrini 2014). In the present paper, we are concerned with the dimensions of quality that affect the probability of grant of a patent application.[1] Therefore, in our context, we understand a high 'quality' patent application as a patent application that has a high probability of grant.

[1] For a patent application to be granted, it must meet the following legal requirements: novelty, inventiveness/non-obviousness, and industrial applicability. Furthermore, the invention must be patentable subject matter, and the patent document must sufficiently disclose the invention.

Various patent quality indicators exist in the literature. One of the most prominent indicators is the number of citations received by a patent. Unfortunately, the count of citations poorly captures grant probability (see Jaffe and de Rassenfosse 2017 for a recent review of the literature). Besides, citation count is not consistently defined across granted and refused patent applications (*i.e.*, the citation arrival process is affected by whether the patent application is eventually granted or not). It is also not consistently defined across local and foreign origins (*i.e.*, examiners may have a preference to cite local prior art, leading to higher citation rates for patents by locals).

Webster *et al.* (2014) have proposed a breakthrough methodological approach to control for the probability of grant of an application. Building on other scholars (Graham *et al.* 2002, Sampat and Amin 2013), the authors exploit 'twin' inventions submitted for patent protection in different jurisdictions. Almost all jurisdictions grant patents for inventions that are new to the world (Correa 2000:58). Thus, in principle, an invention protected by a patent in one country cannot be protected in other countries since the existing patent challenges the worldwide novelty requirement in all the other countries. In practice, patents can be granted for the same invention in different jurisdictions if the applicant of subsequent applications claims the priority of the first application. By using the priority claims, it possible to track the same invention across jurisdictions—so-called 'twin' inventions. One can then use the grant outcome in other patent offices to account for the degree of 'patentability' of the invention. Presumably, an invention granted in all other patent offices is very likely to be granted at the focal office. In contrast, an invention refused in all the other patent offices is also likely to be refused at the focal office.

The set-up of twin inventions is a significant step forward, but it is not sufficient to identify discrimination. Indeed, there might still be systematic differences between foreigners and locals beyond differences in patent 'quality.' For instance, de Rassenfosse and Raiteri (2020) focus on the national treatment principle at the CNIPA. They argue that the quality of the patent agent may systematically differ between foreigners and locals. Controlling for the patent agent and other variables of interest, the authors found no evidence of discrimination against foreigners overall. (However, they did find evidence of discrimination against foreigners in technology areas that the central government considers of ‘strategic importance’.)

Open questions in the search for discrimination

The discussion so far has highlighted the importance of controlling for potential confounding factors in the search for discrimination. Unless we have accounted for all probable sources of heterogeneity between locals and foreigners, we cannot conclude that there is discrimination.

One source of heterogeneity that has not attracted enough attention by scholars arises from the fact that patent prosecution is essentially a negotiation between examiners and applicants, especially at the USPTO. As explained by Lemley and Moore (2004), patent examiners can never effectively reject a patent application. Applicants dissatisfied with the examination decision can argue an unlimited number of times through various mechanisms. This consideration has two potential implications in the present context.

First, it favors wealthier applicants. A 2015 report by the American Intellectual Property Law Association (AIPLA) compiled data from a survey of patent practitioners. According to the report, the median patent agent fees for an original patent application of medium complexity in mechanical engineering reach $9,000, with an interquartile range comprised between $7,500 and $11,000. These fees exclude the filing of amendments or arguments with examiners, which are expensive tasks—a single argument of minimal complexity costs about $2,000 according to the report. Thus, as the patenting process drags on, patenting fees skyrocket. It follows that financial considerations may loom large in the decision to abandon a patent application. Higher odds of getting patents for locals compared to foreigners could be a consequence of the fact that U.S. applicants may be wealthier than foreigners on average.

Second, it is also possible that being on one’s home turf may induce differences in the behavior of firms. Multinational firms usually carry a large proportion of their R&D activities at home (Belderbos *et al.* 2013) such that securing patent protection at home may be particularly important. Besides, analyses of patent data find that the domestic patent office is the office of choice for patenting MNEs, notably for U.S. firms (Criscuolo 2006). Overall, although the United States is a major patenting destination for innovative MNEs of all origins (Beukel and Zhao 2018), we cannot exclude the possibility that U.S. applicants may try harder to have a patent application granted than foreign applicants. Thus, to the extent possible, studies on discrimination in the patent system should account for the effort that applicants put into the prosecution process.

The argument that patent offices may discriminate against foreigners raises the question of whether discrimination is unintentional or intentional. Unintentional discrimination arises when policies, practices, and rules have *disparate impacts* on a specific group of applicants. The discussion so far has focused on unintentional discrimination. By contrast, intentional discrimination relates to *disparate treatment* of a specific group of applicants. Lehmann-Hasemeyer and Streb (2018) offer an interesting account of intentional discrimination in the German state of Wuerttemberg in the 19th century. Discrimination occurred through higher patent fees for foreign states, namely other member states of the German Customs Union. U.S. patent law abides by the national treatment principle such that there is no disparate treatment of foreigners, at least on paper. However, it is also theoretically possible that some individual examiners may not abide by this principle.

A potential mechanism leading to disparate treatment against foreigners is 'ethnocentrism,' namely the tendency to view one's group as centrally important and as superior to other groups (Sumner 1906). Ethnocentrism is a nearly universal syndrome of discriminatory attitudes and behaviors (Sumner 1906, LeVine and Campbell 1972, Bizumic and Duckitt 2012). It manifests itself, among other things, by in-group trust and favoritism (Hammond and Axelrod 2006, Brewer and Gaertner 2001). In the present context, ethnocentrism could lead examiners to exert less effort in their search for prior art for patent applications by ‘trustworthy’ applicants.[2] Similarly, it could also make examiners more lenient towards in-group members for marginal patent applications.

EMPIRICAL APPROACH

Baseline specification

The empirical analysis seeks to model the grant outcome, Y_i, of patent application *i* at the USPTO. The variable Q_i controls for the probability of grant of invention *i* using information from twin patents as explained further below. The variable of interest is the binary indicator F_i. It takes value 1 when the strict majority of inventors are foreign or 0 when they reside in the United States. We use the country of residence of inventors instead of applicants to follow more closely Webster *et al.* (2014).[3] The patent application is filed by firm *f* and patent agent *a* in year *t*. It is examined by patent examiner *e* in art unit u.[4] Our preferred specification is a linear probability model of the form:

$$Y_i = \alpha_t + \alpha_a + \alpha_e + \alpha_u + \beta_1 Q_i + \beta_2 F_i + \boldsymbol{\gamma}\mathbf{X} + \varepsilon_\mathrm{i} \qquad (1)$$

where $\boldsymbol{\gamma}$ is a (1×*K*) vector of coefficients, **X** is a (*K*×1) vector of control variables, and ε_i is a well-behaved disturbance term. The α’s capture fixed effects for application year, patent agent, examiner, and art unit, respectively. One clear advantage of the linear probability model over non-linear models such as probit or logit is that it allows us to control for large dimensional fixed effects—we have about 12,000 individual examiners and more than 19,000 patent agents. We use the user-written Stata command `reghdfe` for that purpose (Correia 2017). However, aware of the limitations of the linear probability

[2] For instance, U.S. examiners could be more suspicious of patent applications by Chinese inventors, in light of recurring debates about the low quality of Chinese patents (Liang 2012, Dang and Motohashi 2015, Boeing and Mueller 2016, Prud’homme and Zhang 2019). Such mental shortcuts are cognitive biases that examiners may not even be aware of.

[3] We have also defined the variable F_i using data on applicants: (i) a dummy variable that takes value 1 when at least one applicant is foreign; and (ii) a dummy variable that takes value 1 when the strict majority of applicant is foreign. Applicant and inventor country of residence exhibit strong correlation. The results are qualitatively and quantitively similar when using alternative definitions of foreignness.

[4] An ‘art unit’ is a working unit inside the USPTO responsible for a cluster of related patent art; it groups examiners by subject matter expertise.

model, we will estimate robust standards errors and limit the interpretation of effect size to the mean effect of the binary variable foreign (F).

Instead of using fixed effects for patent agents, examiners, and art units, an alternative specification involves estimating 'pseudo' fixed effects by computing the average grant rate, excluding patent applications by the focal firm *f*. For instance, the patent agent pseudo fixed effect corresponds to the mean grant rate for all but firm *f*'s patent applications managed by patent agent a. This specification leads to the following linear probability model:

$$Y_i = \alpha_t + a_{-f} + e_{-f} + u_{-f} + \beta_1 Q_i + \beta_2 F_i + \boldsymbol{\gamma}\mathbf{X} + \varepsilon_\mathrm{i} \qquad (2)$$

The pseudo fixed effect approach allows us to estimate additional regression models such as logit and probit. However, they capture heterogeneity in a more limited fashion. For instance, there is one variable per patent agent in specification (1) as opposed to only one variable for all patent agents in specification (2). We will estimate the linear probability model with pseudo fixed effects (model 2) to establish a benchmark against which to compare logit and probit regression results. However, the linear probability model with fixed effects (model 1) remains our preferred specification.

Alternative dependent variables

The binary grant status may be a too crude outcome measure. As explained by Lemley and Moore (2004), patent examiners can never effectively reject a patent application, which makes the patent prosecution process essentially a negotiation between the examiner and the applicant. Once an examiner has reviewed a pending patent application, he/she sends out an 'office action' to the applicant. It takes the form of a letter that indicates the examiner's concerns regarding the claims. The applicant is offered the opportunity to respond to a rejection letter either by arguing that the examiner is incorrect or by amending the claims to overcome the examiner's rejection. The exchange continues until the applicant obtains what he/she wants or abandons. Examiners can send out non-final or final rejections. However, a 'final' rejection is a misnomer because the applicant is afforded the same opportunities as responding to a non-final office action, except that some fees are due if the applicant is amending the claims. It is clear from the discussion that all parties (examiners, assignees, and patent agents) have a strong influence on the outcome of the examination.

This feature of the prosecution process has important implications for our analysis. Should we find that foreigners are less likely than locals to have their patents granted conditional on quality and other covariates, two possible explanations could be put forward. First, examiners could be tougher on foreigners than on locals. Second, foreigners could exert less effort than locals, *i.e.*, try less hard in the face of a rejection. Although these two explanations lead to the same outcome, they have different implications regarding discrimination. The first explanation would suggest disparate treatment, whereas the second explanation would suggest disparate impact.

Therefore, we introduce four more dependent variables in order to shed light on these two potential explanations. First, about assignee effort, we will count the *number of (final and non-final) rejections* (for granted and non-granted patents) to capture how much the assignee has tried to have their patent granted. We will also use the *number of transactions by examiners* and the *number of transactions by applicants* as alternative measures of effort, with more transactions implying more effort (and certainly higher costs).[5] It is important to consider both examiner and applicant transactions jointly, as some examiner transactions can trigger applicant transactions, and vice-versa. Second, regarding examiner effects, we will estimate regression models using *first-action allowance*. A first-action allowance occurs when the patent examiner does not reject any claims in the original patent application and finds that the submission is in order and allowable without amendment. This variable is the closest to the pure action of

[5] The three most frequent categories of transactions initiated by examiners include: “Case Docketed to Examiner in General Art Unit,” “Date Forwarded to Examiner,” and “Non-Final Rejection.” The three most frequent categories of transactions initiated by applicants include: “Information Disclosure Statement Filed (WIDS),” “Information Disclosure Statement Filed (M844),” and “Response after Non-Final Action.” Source: Appendix B to Graham *et al.* (2018) available at https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair.

examiners since no interaction took place between the applicant/patent agent and the examiner at this stage.

DATA

Dataset construction

We constructed our dataset by combining a large number of data sources. However, we have relied primarily on five sources: the PATSTAT database, the USPTO PAIR database, the Google Public Patent database, NamSor application programming interface (API), and NamePrism API. This section provides a brief overview of the data assembly process. Appendix A provides a detailed description of the construction of the dataset, and a GitHub repository contains relevant pieces of software.[6]

The EPO Worldwide Patent Statistical Database (known as PATSTAT) contains detailed bibliographic information on a large number of patent offices worldwide, including the USPTO. The USPTO PAIR database contains information on U.S. published patent applications (Graham *et al.* 2018). We have used the December 2015 version, which contains 9.8 million applications and is available as a public BigQuery dataset hosted on the Google Cloud Platform (GCP). The Google Public Patent database, also available on GCP, contains the claims in plain text for U.S. publications. NamSor API is a commercial classification tool that infers gender, origin, and diaspora information of names using training data collected from various institutional sources.[7] Finally, NamePrism API is a classification software mainly for non-commercial uses that identify the nationality/ethnicity of names using a using 'name embedding' approach trained on large-scale public Twitter data (Ye *et al.* 2017, Ye and Skiena 2019). The NamSor and NamePrism algorithms are among the best methods that we could find to 'guess' ethnic origin. However, one must bear in mind that the limitation of these methods is precisely that they only offer educated guesses about ethnicity.

[6] The GitHub repository is available at the following URL: https://github.com/rezaho/uspto_2019.

[7] Although we are not aware of peer-reviewed validation of NamSor software, we note that it has been used already for academic research (*e.g.*, Hridoy *et al.* 2015, Drechsler *et al.* 2019, Morgan *et al.* 2019).

The identification of twin inventions is a noteworthy feature of our work. We have developed a new type of patent family (we call it ‘twin family’) in which all members are twins. The construction of twin families exploits information on priorities, continuations, and technical relations available in the PATSTAT database. We have allocated all the 96,661,363 applications available in PATSTAT to unique twin families.[8] Appendix B explains the algorithm for building the twin families. We considered patent applications from eight jurisdictions that have a twin application at the USPTO that was filed between the years 2002 and 2012. The eight jurisdictions are the EPO, Japan, mainland China, Korea, Germany, Canada, Australia, and Taiwan. We selected these jurisdictions because they have reasonably accurate grant information in the PATSTAT database, which we need to build the measure of patent application quality.

The final sample contains 1,709,406 patent applications filed between the years 2002 and 2012 and having at least one twin in one of the eight offices.

Variable definitions

We use the following variables in the regression models.

- *Granted (Y)*. Dummy variable that takes value 1 if the U.S. patent application is granted and 0 otherwise, sourced from PATSTAT. Non-granted applications have all received at least one rejection and are either pending or abandoned. Data on abandoned patent applications were available in USPTO PAIR only until 2015, which was not enough for our purpose. We will control for potential differences across cohorts in the mix between pending and abandoned non-granted applications by systematically controlling for application year fixed effects.
- *Foreign (F)*. Dummy variable that takes value 1 if the strict majority of inventors are foreign, as assessed by the country of residence listed in the patent document (and 0 otherwise), sourced from USPTO PAIR.

[8] The full table of twin families is available upon request from the authors.

- *Quality (Q)*. It captures the mean grant rate of (up to eight) applications belonging to the same twin family as the focal U.S. application. This variable is a measure of the extent to which we expect that the patent application should be granted at the USPTO based on the grant outcome at the other patent offices. Information on grant is available in the PATSTAT database.
- *Independent claims*. The number of independent claims listed in the original patent application. Independent claims are the stand-alone features of the invention, and their count captures the scope of the invention (Marco *et al.* 2019). We compute this variable by parsing the full text of the patent document using the Google Public Patent database.
- *Words per claim*. The number of words per independent claim in the patent application. Longer claims are generally narrower and, therefore, protect a slimmer slice of the technology. We compute this variable by parsing the full text of the patent document using the Google Public Patent database.
- *PCT*. Binary variable that takes the value one if the patent application is filed under the Patent Cooperation Treaty (PCT), and 0 otherwise. The PCT is an international agreement that simplifies and lowers the cost of the international patenting process. We sourced the variable from PATSTAT.
- *GDP per capita*. Gross domestic product in current international dollars for the year 2012. For patent documents listing inventors from multiple countries, we took the highest GDP per capita value. The data come from the World Bank Open Data repository. We use this variable as a proxy for financial resources.
- *Small entity*. Dummy variable that takes the value 1 if the patent application was subject to the small-entity reduction fees, and 0 otherwise, sourced from USPTO PAIR. We use this variable as a proxy for financial resources.
- *Portfolio size*. The number of patent applications filed by the assignee in the 5-year before the focal patent application, sourced from PATSTAT. We use this variable as a proxy for financial resources and applicant experience.

- Pseudo fixed effects (PFE) for patent agents, examiners, and art units correspond to the average grant rate for patent applications by the relevant entity (patent agent, examiner, or art unit), excluding applications for the focal assignee. Consider the following example. Let us assume that there are only two art units: ‘A’ and ‘B,’ and three firms: ‘1’, ‘2’, and ‘3.’ In order to compute the PFE for art unit ‘A’ for firm ‘1,’ we have only considered applications that were examined in art unit A by firms ‘2’ and ‘3.’ This approach ensures that the PFE variables are exogenous to the applicant (and a fortiori to the focal patent application). Data on patent agents, examiners, and art units come from the USPTO PAIR database.

We implement two ways of testing for the presence of ethnocentrism in patent examination. First, we identify groups of inventors who, *a priori,* could be facing discrimination. Observers have reported signs of sinophobia and islamophobia in the U.S. population (*e.g.*, Lyman 2000, Ogan *et al.* 2014). Thus, if there is intentional discrimination, these two groups are likely to be potential targets. We also consider a third group, inventors of Japanese origin, because it represents the largest group of foreign inventors in our sample. A large body of work, predominantly in labor economics, has established that a person’s name could vehiculate stereotypes that would lead to discrimination (*e.g.*, Bertrand and Mullainathan 2004, Carlsson and Rooth 2007, Kaas and Manger 2012). Therefore, we rely on name-based ethnic detection algorithms to allocate inventors into the three groups. We have used the NamSor API to identify Japanese and Chinese inventors. We relied on triples of (first name, last name, country code) when the country code information is available or pairs (first name, last name) when it was not. The identification of Muslim inventors relies on the NamePrism API. We fed pairs of (first name, last name) to the API and selected the nationality that was returned. We then manually associated inventors with the following ‘nationalities’ (regions of origin) as likely to be Muslim: Arabian Peninsula, Maghreb, Nubian, Pakistanis-Bangladesh, Pakistanis-Pakistan, Persian, Turkic-CentralAsian, and Turkic-Turkey.

Second, one could argue that the cultural distance between the examiner and the inventor is a more appropriate way of capturing potential discrimination than focusing on specific groups of inventors.

We measure distance in two ways. First, we have run both inventor and examiner names through the NamSor API. The variable '*Same country of origin*' takes value 1 when the country of origin for the pair (first name, last name) of the examiner matches the country of origin for the pair (first name, last name) of at least one inventor listed in the patent document. Although U.S. examiners must be U.S. citizens or U.S. permanent residents, their name may be associated with a different country of origin than the United States. Second, we rely on a more fine-grained indicator to measure cultural proximity that a binary variable. Social psychologist Geert Hofstede has found that differences in national cultures vary substantially along six dimensions (Hofstede 1980, Hofstede *et al.* 2010). These dimensions are labeled individualism, tolerance of power distance, masculinity, uncertainty avoidance, long-term orientation, and indulgence. Following Konara and Mohr (2019), the variable '*Cultural proximity*' captures the standardized Euclidean distance between the (country of residence of) the examiner and the inventor using Hofstede's six dimensions.[9] In case inventors come from multiple countries, we compute the score using the culturally-closest country.

Descriptive statistics

Table 1 and Table 2 present an overview of the data. The overall grant rate at the USPTO in our sample is 66 percent, but 75 percent for domestic inventors and 64 percent for foreign inventors. About 14 percent of patents are granted at first office action. Patent applications have an average of 1.82 rejections and are subject to 6.57 applicant transactions and 13.65 examiner transactions. Foreigners file about 87 percent of patent applications in the sample. Chinese inventors account for 4 percent of patent applications in our sample, and Japanese inventors for 33 percent. A mere 0.4 percent of patent applications arise from inventors located in countries where Islam is the main religion. The average quality score is 52 percent, suggesting that about half of the twins are granted in the other jurisdictions in which they are filed. This variable may be a suitable candidate to predict the likelihood of a grant at the USPTO, as indicated by the strong correlation with the variable *Granted* (correlation coefficient of 0.33, not reported). The GDP per

[9] The data are available at https://geerthofstede.com/research-and-vsm/dimension-data-matrix/.

capita varies from a minimum of 365 PPP for Mauritania to 126,618 PPP for Macao SAR, China. The size of the patent portfolio ranges from a minimum of 0 (*i.e.*, no previous patent experience) to a maximum of 44,627 for a patent application by IBM Corporation. About 13 percent of assignees are small entities (as indicated by the payment of reduced fees), and the majority of patent applications by small entities are from foreign origin (about 87 percent, not reported).

Turning to individual effects, Table 2 indicates that the sample contains 19,310 patent agents having prosecuted an average of 80.84 patent applications. The largest patent agent in the sample is Oblon, McClelland, Maier & Neustadt, L.L.P., with about 54,000 patent applications. More than 12,000 examiners belonging to 709 art units examined the patent applications in the sample.

Table 1. Descriptive statistics for regression variables

	N	Min	Mean	Max	Std. Dev.
Granted* (*Y*)	1,56,1173	0	0.66	1	-
… domestic*	195,794	0	0.75	1	-
… foreign*	1,365,379	0	0.65	1	-
Granted at first office action*	1,561,176	0	0.14	1	-
No. of rejections	1,56,1173	0	1.82	24	1.67
No. of applicant transactions	1,56,1173	0	6.57	800	4.88
No. of examiner transactions	1,56,1173	0	13.65	1312	9.26
Foreign* (*F*)	1,56,1173	0	0.87	1	-
Foreign & Chinese*	1,56,1173	0	0.04	1	-
Foreign & Japanese*	1,56,1173	0	0.33	1	-
Foreign & Muslim*	1,56,1173	0	0.004	1	-
Same country of origin*	1,56,1173	0	0.10	1	-
Cultural distance	1,510,511	0	1.51	2.48	0.83
Quality (*Q*)	1,56,1173	0	0.52	1	0.42
Independent claims	1,500,038	0	4.80	38,649	33.67
Words per claim	1,497,789	1	108	9,511	83.62
PCT*	1,561,173	0	0.32	1	-
log(GDP per capita)	1,433,573	6.64	10.56	11.75	0.30
log(Portfolio size)	1,414,952	0	4.58	10.44	3.12
Small entity*	1,56,1173	0	0.13	1	-

Notes: ‘*’ indicates a dummy variable.

Table 2. Descriptive statistics for individual effects

	Total number of applications	Number of unique observations	Minimum number of applications	Average number of applications	Maximum number of applications
Patent agent	1,561,085	19,310	1	80.84	53,661

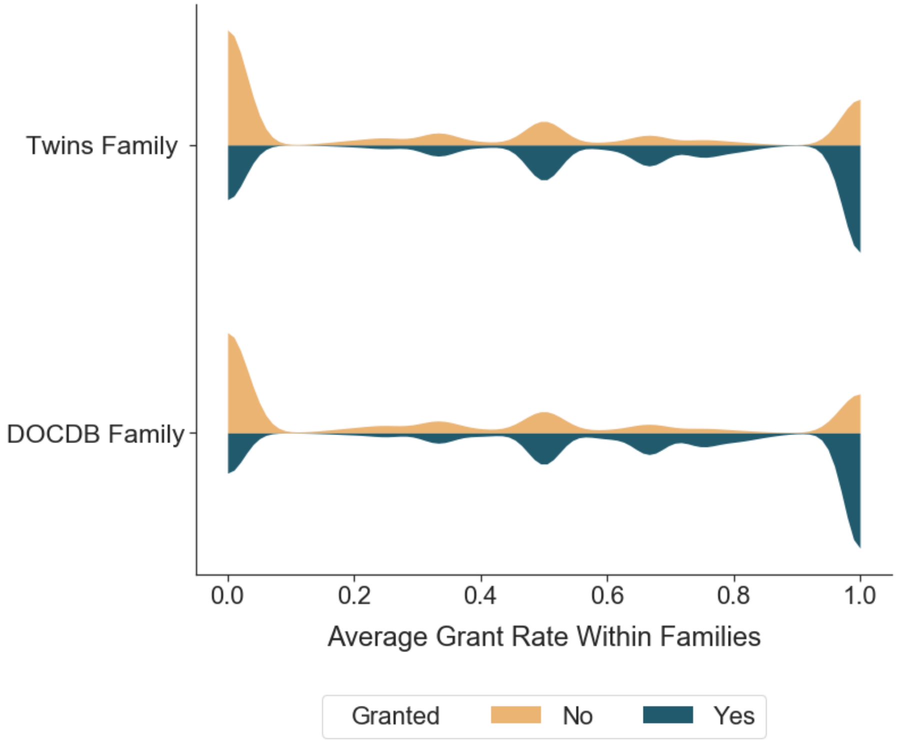
Examiner	1,561,176	12,306	1	149.59	1,471
Art unit	1,561,173	709	1	2604.48	14,620
Filing Year	1,561,173	11	148,111	162,764	180,978

The next figures illustrate some key dimensions of the data. They provide a first glimpse into potential reasons for the difference in grant rates between foreigners and locals documented in Table 1. However, we will investigate these reasons in detail in the next section.

As explained above, the ‘quality’ variable is the mean grant rate of a set of twin inventions submitted to different offices. The twin family offers a more stringent definition of family than existing definitions (see Martínez 2011 for an overview of the various definitions). The family definition that is the closest in spirit to our approach is the DOCDB simple patent family, which is constructed by EPO examiners.[10] We define the twin family as a collection of patent applications that are considered to cover a single invention (in the sense that the technical content covered by the applications is considered to be identical), and in which family members all share the same priorities. Figure 1 compares the distributions of the average grant rate among twin families (upper panel) and DOCDB families (lower panel), for families for which the U.S. member is granted (light color) and for which it is not (dark color). The distributions look very similar across family definitions, suggesting that the twin family is representative of the overall population.

[10] DOCDB families include patent documents that share identical ‘priority pictures’ (that is, priorities adding new technical content after having excluded redundant priorities via expert control), see Martínez (2011).

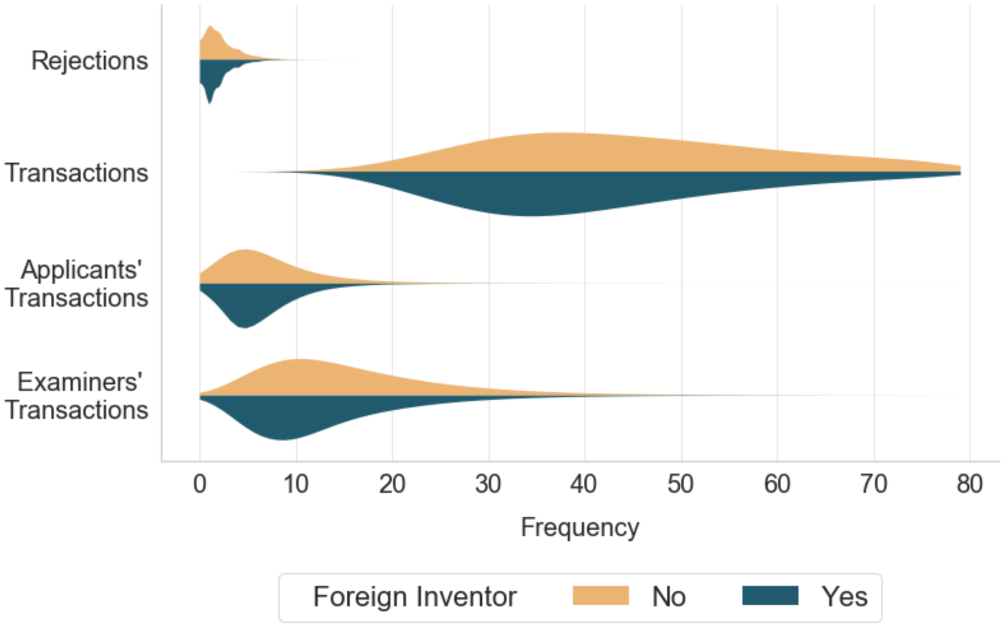
Figure 1. Kernel density plot of average grant rate among families, by type of family and grant status at the USPTO.



Notes: Data for ‘Twins Family’ and ‘DOCDB Family’ limited to eight jurisdictions and for the study period 2002–2012. Random sample of 1 million observations for each family type. See the main text for the definitions of families.

Figure 2 provides an overview of different types of office actions for patents by foreigners and by locals. There is consistently more mass to the left of the distributions for foreigners compared to locals. Patent applications by foreigners receive fewer rejections (the median number is 1 for foreigners and 2 for locals, not reported). They are also associated with a lower number of transactions (the median number is 39 for foreigners and 43 for locals, not reported). The difference in the number of transactions between foreigners and locals is driven primarily by examiners (the median number of examiner transactions is 11 for foreigners and 14 for locals, not reported). These figures suggest that foreigners may abandon sooner than locals (in case the application is abandoned) or may sail through the patenting process more smoothly (in case the application is granted).

Figure 2. Distributions of office actions for patent applications by foreigners and locals

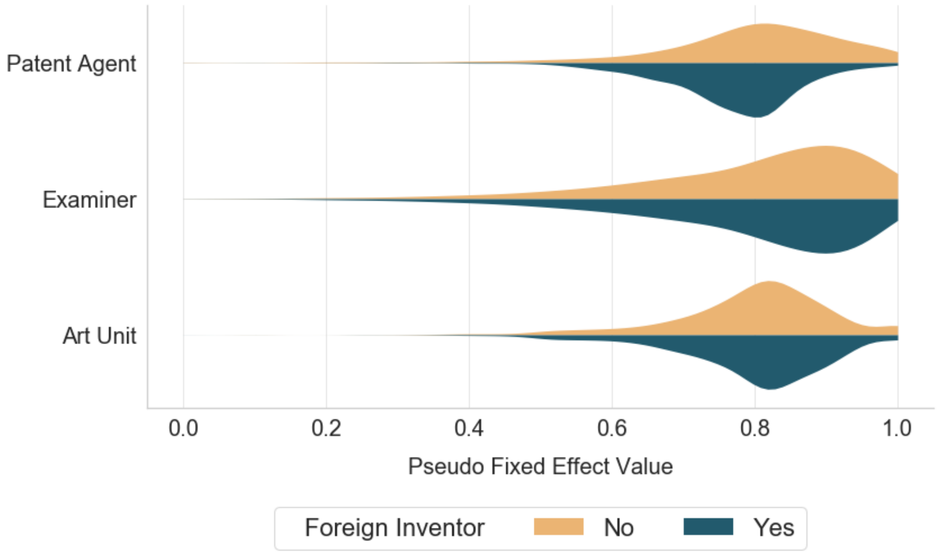


Notes: Kernel distributions reported. Data for the study period 2002–2012 for U.S. applications having at least one twin in any of the eight jurisdictions considered. Random sample of 1 million observations for each office action type.

In Figure 3, we use the pseudo fixed effects to investigate whether there are systematic differences in patent agents, examiners, and art units between foreigners and locals. It seems that patent agents of foreigners have lower success rates than patent agents of locals.[11] The interquartile range of the patent agent pseudo fixed effect is 0.75–0.88 for locals and 0.72–0.82 for foreigners (not reported). There are no noticeable differences between foreigners and locals regarding the examiner and art unit pseudo fixed effects.

[11] We refrain from interpreting the average grant rate of patent agents as a measure of patent agents' intrinsic quality. First, systematic differences in applicant budgets may explain differences in patent agent effort, and therefore grant rates. Second, patent agents may also differ systematically in the type of inventions they prosecute.

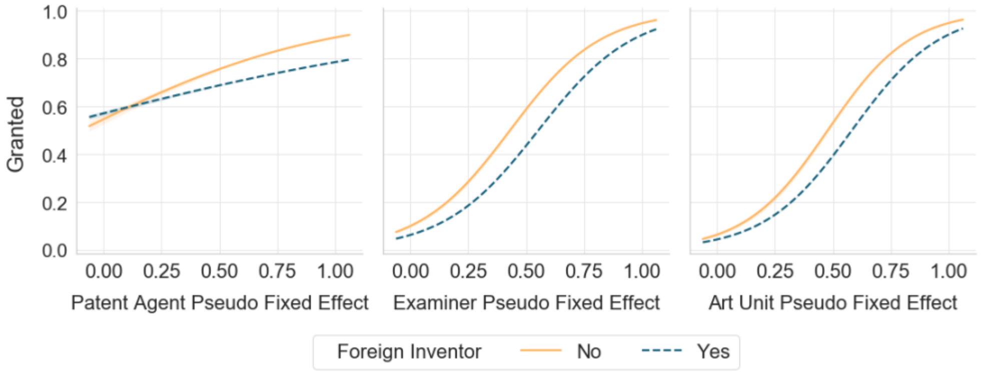
Figure 3. Distributions of selected pseudo fixed effects for locals versus foreigners.



Notes: Kernel distributions reported. Data for the study period 2002–2012 for U.S. applications having at least one twin in any of the eight jurisdictions considered. Random sample of 100,000 million observations for each effect type.

Figure 4 provides tentative evidence that: i) there are significant patent agent, examiner, and art unit effects in terms of grant outcome; and ii) that foreigners have a lower probability of grant than locals for all levels of the pseudo fixed effects. For instance, for patent agents of average score (say with a value at 0.80), foreigners have an average grant probability of 75 percent compared to 84 percent for locals, as indicated in the left panel of Figure 4. We also know from Figure 3 that foreigners tend to have agents with lower scores than locals, which further increases the spread in grant probability between foreigners and locals.

Figure 4. Logistic regression plot of granted at USPTO as a function of pseudo fixed effects for foreigners versus locals.



ECONOMETRIC RESULTS

We start with a simple specification and gradually control for features of the patent applications in Table 3. Column (1) suggests that U.S. patent applications by foreigners are 9.6 percentage points less likely to be granted than patent applications by locals. However, as discussed in Section 2, there may be systematic differences between patent applications by locals and by foreigners such that this spread cannot be taken as evidence of discrimination. Accounting for the grant outcome of the twin applications in foreign jurisdictions in column (2) increases the apparent bias against foreigners slightly. The increase in foreign bias suggests that foreigners submit higher quality patent applications on average. Next, controlling for the scope of the patent application, as well as its filing route, in column (3) leaves the coefficient of interest unchanged.

As shown in Figure 4, there are significant differences in the average grant rate of patent agents, examiners, and art units. Accounting for these differences using a set of fixed effects contributes to explaining the likelihood of grant, as indicated by the doubling of the adjusted R_2 from column (3) to column (4). The 'bias' against foreigners decreases by 4.2 percentage points to 6.9 percentage points and remains significant at the 0.1-percent probability threshold.

The next three regression models exploit the pseudo fixed effects. Column (5) reports the OLS estimates as a benchmark, whereas column (6) reports the probit estimates and column (7) the logit estimates. The three methods produce sensibly similar coefficients, and the foreign bias reaches about 12 percentage points. Our preferred specification is the binary fixed-effect linear probability model (column 4) because it is a richer specification.

Table 3. Baseline specification controlling for 'quality' and fixed effects or pseudo fixed effects (PFE).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Econometric Method:	OLS	OLS	OLS	OLS	OLS	Probit	Logit
Foreign (*F*)	-0.096**	-0.112**	-0.111**	-0.069**	-0.118**	-0.125**	-0.125**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Quality (*Q*)		0.392**	0.383**	0.345**	0.349**	0.320**	0.322**
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
log(Independent claims)			0.015**	0.021**	0.027**	0.027**	0.027**
			(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
log(Words per claim)			0.066**	0.050**	0.043**	0.042**	0.041**
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PCT			0.012**	0.028**	0.041**	0.036**	0.033**
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Patent agent PFE					0.119**	0.126**	0.120**
					(0.003)	(0.003)	(0.003)
Examiner PFE					0.786**	0.701**	0.691**
					(0.003)	(0.002)	(0.002)
Art unit PFE					0.167**	0.159**	0.158**
					(0.004)	(0.004)	(0.004)
Patent agent FE	No	No	No	Yes	No	No	No
Examiner FE	No	No	No	Yes	No	No	No
Art unit FE	No	No	No	Yes	No	No	No
Application year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R2	0.020	0.142	0.152	0.319	0.254	-	-

Notes: The dependent variable is the dummy '*Granted*';
N = 1,344,867;
Marginal effects at mean reported in columns (6)–(7);
Constant term included but not reported;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

Next, we investigate whether differences in financial resources and experience help in explaining the apparent foreign bias. The regression model presented in column (1) of Table 4 is similar to that of column (4) of Table 3. We use it to assess the sensitivity of the results to sample composition (control variables are not available for all observations). In columns (2) to (4), we add three variables correlated with financial resources and experience. It is admittedly challenging to tease out experience from financial resources—wealthier applicants also have more opportunities to accumulate experience—and we refrain from doing so. Firms from richer countries (column 2) and with a larger portfolio of patent applications (column 3) are more likely to have their patents granted. Interestingly, small entities are significantly less likely to have their patent applications granted (column 4). However, the negative effect disappears once we control for the size of the patent portfolio (not reported).

Controlling for all the variables jointly leads to a reduction in the apparent bias against foreigners, which settles at 7.7 percentage points in column (5). To some extent, these results suggest unequal access to the U.S. patent system: accumulated experience and financial resources help in getting a patent—this is one reason why foreigners (and small firms) have a relatively low success rate at the USPTO.

Table 4. Controlling for financial resources and experience

	(1)	(2)	(3)	(4)	(5)
Foreign (*F*)	-0.085**	-0.074**	-0.086**	-0.086**	-0.077**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Quality (*Q*)	0.351**	0.353**	0.340**	0.351**	0.341**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
log(Independent claims)	0.017**	0.017**	0.014**	0.017**	0.014**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
log(Words per claim)	0.047**	0.047**	0.041**	0.047**	0.041**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PCT	0.029**	0.029**	0.060**	0.031**	0.059**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
log(GDP per capita)		0.035**			0.028**
		(0.001)			(0.002)
log(Portfolio size)			0.034**		0.034**
			(0.000)		(0.000)
Small entity				-0.051**	0.023**
				(0.001)	(0.002)
Patent agent FE	Yes	Yes	Yes	Yes	Yes
Examiner FE	Yes	Yes	Yes	Yes	Yes
Art unit FE	Yes	Yes	Yes	Yes	Yes
Application year FE	Yes	Yes	Yes	Yes	Yes
Adjusted R_2	0.322	0.322	0.350	0.323	0.351

Notes: The dependent variable is the dummy ‘*Granted*’;
Econometric method is OLS;
N = 1,240,980;
Constant term included but not reported;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

Do foreigners put in less effort than domestic inventors?

As explained in the ‘Background’ section, foreigners may make less effort than locals to have their patent applications granted, which may explain the difference in grant rate partly. In the next three tables, we investigate whether we find some empirical support for this explanation. Table 5 estimates the determinants of the number of rejections for the subsample of non-granted applications (column 1) and the subsample of granted applications (column 2). Overall, foreigners interact less with examiners than domestic applicants, both for non-granted applications and for granted patents (and so do small entities). The negative coefficient associated with the variable *Foreign* in column (1) suggests that foreigners try less (*i.e.*, abandon faster) than locals. The negative coefficient in column (2) suggests that foreigners receive, on average, 0.14 fewer rejections before their patent application is granted. A lower number of rejections could reflect the fact that foreigners may better respond to examiner requests (*i.e.*, they may concede more).

Table 5. Determinants of the number of rejections

	(1)	(2)
Dependent variable:	Number of rejections	
Sample:	Non-granted applications	Granted applications
Foreign (*F*)	-0.538**	-0.143**
	(0.015)	(0.006)
Quality (*Q*)	1.037**	-0.003
	(0.008)	(0.004)
log(Independent claims)	-0.054**	-0.073**
	(0.004)	(0.002)
log(Words per claim)	-0.213**	-0.360**
	(0.005)	(0.002)
PCT	-0.355**	-0.213**
	(0.007)	(0.004)
log(GDP per capita)	0.082**	0.095**
	(0.010)	(0.006)
log(Portfolio size)	0.011**	-0.015**
	(0.001)	(0.001)
Small entity	-0.308**	-0.174**
	(0.009)	(0.006)
Patent agent FE	Yes	Yes
Examiner FE	Yes	Yes
Art unit FE	Yes	Yes
Application year FE	Yes	Yes
Adjusted R_2	0.325	0.323
N	364,178	874,190

Notes: Econometric method is OLS;
Constant term included but not reported;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

An alternative measure of effort is the count of the number of ‘transactions’ having occurred during patent prosecution—with more transactions implying more effort (and certainly higher costs). As explained in the ‘Empirical Approach’ section, transactions can be induced either by examiners or by applicants. Sometimes, examiner transactions can trigger applicant transactions and vice-versa. Table 6 reports the results of seemingly unrelated regression (SUR) models for the joint estimation of the determinants of the number of examiner and applicant transactions. Columns (1a)–(1b) focus on the sample of all patent applications, columns (2a)–(2b) on the subsample of non-granted patents, and columns (3a)–(3b) on the subsample of granted patents. The coefficients associated with the variable *Foreign* are always negative, suggesting that foreigners put less effort into the prosecution process (on their own volition or not).

Table 6. Seemingly Unrelated Regression Models

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Dependent variable:	EXA	APP	EXA	APP	EXA	APP
Sample:	All observations		Only non-granted		Only granted	
Foreign (*F*)	-2.846**	-1.277**	-4.422**	-1.946**	-1.959**	-0.959**
	(0.026)	(0.014)	(0.055)	(0.028)	(0.028)	(0.016)
Quality (*Q*)	3.137**	1.176**	6.284**	2.217**	0.328**	0.316**
	(0.019)	(0.010)	(0.039)	(0.020)	(0.023)	(0.013)
log(Independent claims)	0.752**	0.494**	0.788**	0.506**	0.668**	0.478**
	(0.011)	(0.006)	(0.021)	(0.010)	(0.013)	(0.008)
log(Words per claim)	-0.713**	-0.204**	-0.376**	-0.132**	-0.969**	-0.261**
	(0.012)	(0.007)	(0.022)	(0.011)	(0.014)	(0.008)
PCT	-1.329**	0.416**	-1.911**	0.306**	-1.298**	0.390**
	(0.018)	(0.010)	(0.035)	(0.018)	(0.020)	(0.012)
log(GDP per capita)	0.589**	0.371**	0.629**	0.406**	0.405**	0.313**
	(0.028)	(0.015)	(0.051)	(0.026)	(0.032)	(0.019)
log(Portfolio size)	0.119**	0.035**	0.160**	0.053**	0.061**	0.009**
	(0.003)	(0.002)	(0.006)	(0.003)	(0.003)	(0.002)
Small entity	-1.333**	-0.869**	-1.904**	-0.987**	-1.079**	-0.846**
	(0.028)	(0.015)	(0.046)	(0.023)	(0.034)	(0.020)
Patent agent PFE	0.925**	1.239**	2.983**	2.083**	-0.143	0.828**
	(0.071)	(0.039)	(0.134)	(0.068)	(0.080)	(0.047)
Examiner PFE	-12.664**	-4.205**	-8.366**	-2.787**	-19.937**	-6.408**
	(0.056)	(0.031)	(0.094)	(0.048)	(0.071)	(0.041)
Art unit PFE	-3.500**	-1.940**	-2.346**	-1.639**	-5.055**	-2.400**
	(0.084)	(0.046)	(0.141)	(0.071)	(0.100)	(0.059)
Application year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,244,048		366,472		877576	
R$_2$	0.143		0.173		0.196	

Notes: Estimates obtained from seemingly unrelated regression (SUR) models;
'EXA': Number of examiner transactions, 'APP': Number of applicant transactions;
Constant term included but not reported;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

Having established that foreigners may tend to put less effort into their patent applications than locals, we next sought to assess how much of the apparent bias can be accounted for by differences in effort. The regression models in Table 7 use the number of rejections and the total number of transactions as proxies for effort. When we control for both variables, the spread drops to 3.9 percentage points (column 3). Taken together, these results suggest that part of the apparent bias that we observe is because foreign applicants put less effort in the patent prosecution process.

Table 7. Accounting for applicant effort

	(1)	(2)	(3)
Dependent variable:	Granted (*Y*)		
Foreign (*F*)	-0.083**	-0.050**	-0.039**
	(0.001)	(0.001)	(0.001)
Quality (*Q*)	0.346**	0.288**	0.256**
	(0.001)	(0.001)	(0.001)
log(Independent claims)	0.011**	0.004**	-0.008**
	(0.001)	(0.001)	(0.000)
log(Words per claim)	0.037**	0.052**	0.023**
	(0.001)	(0.001)	(0.001)
PCT	0.062**	0.063**	0.024**
	(0.001)	(0.001)	(0.001)
log(GDP per capita)	0.026**	0.008**	0.009**
	(0.002)	(0.001)	(0.001)
log(Portfolio size)	0.036**	0.035**	0.027**
	(0.000)	(0.000)	(0.000)
Small entity	0.025**	0.043**	0.024**
	(0.002)	(0.001)	(0.001)
Number of rejections	-0.011**		-0.126**
	(0.000)		(0.000)
Number of transactions		0.007**	0.014**
		(0.000)	(0.000)
Patent agent FE	Yes	Yes	Yes
Examiner FE	Yes	Yes	Yes
Art unit FE	Yes	Yes	Yes
Application year FE	Yes	Yes	Yes
R2	0.345	0.404	0.484

Notes: The dependent variable is the dummy ‘*Granted*’;
Econometric method is OLS;
N = 1,240,999;
Constant term included but not reported;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

Can we find traces of ethnocentrism?

The grant outcome is the result of a complex negotiation process between examiners, patent agents, and applicants. In order to test whether examiners are biased against a specific group of applicants, we consider another outcome variable in Table 8, namely a dummy capturing whether the patent application was granted at first office action. Since a grant at first office action is not affected by applicant and patent agent behavior (other than through the drafting of the application), this variable offers a cleaner test of potential discrimination.

The results in column (1) suggest that foreign inventors are 0.7 percentage points more likely than domestic inventors to be granted a patent in the smoothest possible manner. Although the effect is positive, suggesting positive discrimination, it is admittedly small in magnitude. The specification in column (2) breaks down the effect by specific groups of inventors. We find no discrimination against Chinese, Japanese, and Muslim inventors. If anything, we find evidence of positive discrimination, especially concerning Japanese inventors. One could explain the higher grant rate for patent applications by Japanese inventors by the fact that patents at the JPO make narrow claims. Since many of the U.S. applications by Japanese inventors would be second filings claiming priority from the Japanese patent, the U.S. equivalent makes presumably narrow claims as well—it is thus more likely to be granted smoothly compared to patents that make broader claims.

The last two columns control for measures of proximity between examiners and inventors. When their names suggest that they come from the same origin country (*e.g.*, a U.S. examiner with an Indian name matched to an inventor with an Indian name), the probability of a grant at first office action decreases by 0.5 percentage points (column 3). We obtain a similar finding with the measure of cultural distance in column (4)—this time treating U.S. examiners as sharing U.S. values on Hofstede's scale. Inventors that are culturally closer to the United States have a lower probability of a grant at first office action. However, note that the coefficient associated with the variable *Foreign* has become insignificant. This result suggests that the coefficient associated with the variable *Cultural distance* is driven primarily by U.S. inventors being less likely than foreigners to receive a grant at first office action.

Overall, the estimates suggest that foreigners have a *higher* (though modest) likelihood of grant at first office action. Therefore, we find no trace of intentional discrimination from U.S. examiners against foreign inventors.

Table 8. Determinants of granted at first office action

	(1)	(2)	(3)	(4)
Dependent variable:	Granted at First Office Action			
Foreign (*F*)	0.007** (0.001)	0.004** (0.001)	0.007** (0.001)	0.003 (0.001)
Foreign & from China		0.007** (0.002)		
Foreign & from Japan		0.014** (0.001)		
Foreign & Muslim		-0.003 (0.005)		
Same country of origin			-0.005** (0.001)	
Cultural distance				0.004** (0.001)
Quality (*Q*)	0.058** (0.001)	0.058** (0.001)	0.058** (0.001)	0.058** (0.001)
Small entity	0.015** (0.001)	0.017** (0.001)	0.015** (0.001)	0.016** (0.001)
Patent agent FE	Yes	Yes	Yes	Yes
Examiner FE	Yes	Yes	Yes	Yes
Art unit FE	Yes	Yes	Yes	Yes
Application year FE	Yes	Yes	Yes	Yes
Adjusted R_2	0.194	0.194	0.194	0.194

Notes: Econometric method is OLS;
N = 1,240,980;
The regressions control for the following variables (not reported): log(Independent claims), log(Words per claim), PCT, log(GDP per capita), log(Portfolio size), constant term;
Robust Standard errors in parentheses;
* $p < 0.01$, ** $p < 0.001$.

DISCUSSION

Motivated by emerging scholarly evidence that patent offices around the world may violate the national treatment principle, this paper studies the facts using USPTO data. An analysis of the examination outcome of about 1.5 million U.S. patent applications filed in the years 2002–2012 suggests that applications of foreign origin are ten percentage points less likely to be allowed than domestic applications. Although one may interpret this finding as evidence of disparate treatment of foreigners, it turns out that the difference in grant rates can be explained in large part by three sources of heterogeneity: patent agents, financial resources, and effort.

The main results are as follows. First, the data suggest that patent agents exhibit a high level of heterogeneity in how likely they are to get patents granted, and that foreigners seem to select patent agents with lower grant rates. This difference in the success rate of patent agents between foreigners and locals could be due to the dedication of a more limited budget by foreigners or to a difference in mandate. For instance, some patent agents for international applicants may primarily conduct translation services and offer limited bargaining with the patent examiners over the technical aspects of an invention. Shedding more light on this issue would require data on patent agent fees and their mandate, which we do not have. Nevertheless, we observe significant differences in terms of patent agents between foreigners and locals, and these differences help understand the lower grant rate for foreign applicants compared to locals.

Second, the grant probability increases with the wealth of the origin country and the size of the applicant patent portfolio. Controlling for these factors further reduces the spread in grant probability between foreigners and locals. Third, it seems that foreigners generate a lower number of transactions in the course of the prosecution process than locals, suggesting that they may be putting less effort. Adding controls for proxies of effort to the regression models reduces the spread in grant rates to significantly lower levels. Overall, we do not find evidence of unfair treatment of foreigners. On the contrary, an analysis of allowances at first office action reveals a positive, if negligible, advantage for foreigners.

Overall, the USPTO seems to uphold the national treatment principle. The story that emerges from the analysis is more about ‘disparate impact’ than ‘disparate treatment.’ All of the three channels that we have identified directly relate to the financial resources of the applicant. Given the disparity in the fees asked by patent agents (AIPLA 2015), better patent agents are also presumably more expensive.

Furthermore, having a large patent portfolio signals a large patent budget. Finally, responding to rejections and interacting with the patent office is costly, such that applicants with deeper pockets can afford to fight longer to have their patent applications granted. In that respect, it is telling that empirical patterns between foreigners and small entities are similar in many specifications.

The finding that financial resources matter has implications that go beyond the issue of the national treatment principle. It suggests, first and foremost, that applicants do not have equal access to the patent system, which may reinforce the position of large firms vis-à-vis smaller ones. Larger and wealthier firms, who also presumably hold stronger market positions, can afford to hire the best agents and take the time to wear down examiners. Startups and SMEs, on the other hand, cannot play this game. Thus, the patent system may maintain an uneven playing field instead of leveling it. The finding also has implications for the discussion on patent quality (Jaffe and Lerner 2004, Lemley and Shapiro 2005, Bessen and Meurer 2008). The empirical analysis estimates the determinants of a U.S. grant, controlling for the average grant rate of the same invention at other offices. Therefore, one could argue that factors that increase the grant probability at the USPTO (holding constant the average grant rate at other offices) lead to a deterioration of patent quality at the USPTO (compared to other offices). In this regard, the fact that U.S. patent law allows to argue indefinitely with examiners can be seen as one root cause for the relatively low quality of U.S. patents compared to other jurisdictions (de Rassenfosse *et al.* 2019).

We conclude with one policy message. The fact that examiners cannot effectively reject a patent application opens the doors to endless discussions and negotiations between the examiners and the patent agent/applicant, and eventually to a patent allowance. It is predominantly wealthier applicants that benefit from this system. Giving examiners the power of (truly) rejecting a patent application may be one solution to level the playing field between foreigners and locals, but also between large and small firms.

REFERENCES

American Intellectual Property Law Association. 2015. *2015 Report of the Economic Survey*. Arlington, Virginia.

Belderbos, R., Leten, B., & Suzuki, S. 2013. How global is R&D? Firm-level determinants of home-country bias in R&D. *Journal of International Business Studies*, 44(8): 765–786.

Bertrand, M., & Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013.

Bessen, J. E., & Meurer, M. J. 2008. *Patent failure: How judges, bureaucrats, and lawyers put innovators at risk*. Princeton University Press.

Beukel, K., & Zhao, M. 2018. IP litigation is local, but those who litigate are global. *Journal of International Business Policy*, 1(1–2), 53–70.

Bizumic, B., & Duckitt, J. 2012. What is and is not ethnocentrism? A conceptual analysis and political implications. *Political Psychology*, 33(6), 887–909.

Boeing, P., & Mueller, E. 2016. Measuring patent quality in cross-country comparison. *Economics Letters*, *149*, 145–147.

Brander, J. A., Cui, V., & Vertinsky, I. 2017. China and intellectual property rights: A challenge to the rule of law. *Journal of International Business Studies*, 48(7): 908–921.

Brewer, M. B., & Gaertner, S. L. 2001. Toward reduction of prejudice: Intergroup contact and social categorization. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp. 451–472). Malden, MA: Blackwell.

Carlsson, M., & Rooth, D. O. 2007. Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics*, 14(4): 716–729.

Correa, C. M. 2000. *Intellectual property rights, the WTO and developing countries: the TRIPS agreement and policy options*. Zed books.

Correia, S. 2017. REGHDFE: Stata module for linear and instrumental-variable/gmm regression absorbing multiple levels of fixed effects. Statistical Software Components s457874. Available at: https://ideas.repec.org/c/boc/bocode/s457874.html.

Criscuolo, P. 2006. The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics*, 66(1): 23–41.

Dang, J., & Motohashi, K. 2015. Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality. *China Economic Review*, 35: 137–155.

de Rassenfosse, G., Dernis, H., & Boedt, G. 2014. An introduction to the Patstat database with example queries. *Australian Economic Review*, 47(3): 395–408.

de Rassenfosse, G., Griffiths, W. E., Jaffe, A. B., & Webster, E. 2019. Low-quality patents in the eye of the beholder: Evidence from multiple examiners. *National Bureau of Economic Research* WP No. 22244.

de Rassenfosse, G., Jensen, P. H., Julius, T., Palangkaraya, A., Webster, E. 2019. Are Foreigners Treated Equally under the Trade-Related Aspects of Intellectual Property Rights Agreement? *Journal of Law & Economics*, 62(4): 663–685.

de Rassenfosse, G., & Raiteri, E., 2020. Technology Protectionism and the Patent System: Strategic Technologies in China. *Journal of Industrial Economics*, forthcoming.

Drechsler, J., Bachmann, J. T., & Engelen, A. 2019. The effect of immigrants in the founding team on the international attention of new ventures. *Journal of International Entrepreneurship*, 17: 305–339.

Graham, S. J., Hall, B. H., Harhoff, D., & Mowery, D. C. 2002. Post-issue patent ‘quality control’: A comparative study of US patent re-examinations and European patent oppositions. *National Bureau of Economic Research Working Paper* No. 8807.

Graham, S. J., Marco, A. C., & Miller, R. 2018. The USPTO patent examination research dataset: A window on patent processing. *Journal of Economics & Management Strategy*, 27(3): 554–578.

Guerrini, C. J. 2014. Defining patent quality. *Fordham Law Review*, 82(6): 3091–3141.

Hammond, R. A., & Axelrod, R. 2006. The evolution of ethnocentrism. Journal of Conflict Resolution, 50(6), 926–936.

Harhoff, D., & Wagner, S., 2009. The duration of patent examination at the European Patent Office. *Management Science*, 55(12): 1969–1984.

Harris, D. P. 2009. The honeymoon is over: Evaluating the United States’ WTO intellectual property complaint against China. *Fordham International Law Journal*, 32: 2008–2076.

Helfgott, S. 1990. Cultural differences between the U.S. and Japanese patent systems. *Journal of the Patent & Trademark Office Society*, 72: 231–238.

Hofstede, G. 1980. *Culture’s consequences: International differences in work-related values*. Beverly Hills: Sage Publications.

Hofstede, G., Hofstede, G. J., & Minkov, M. 2010. *Cultures and Organizations: Software of the Mind*, 3rd edition.

Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. 2015. Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2(1): 8.

Ivus, O. 2015. Does stronger patent protection increase export variety? Evidence from US product-level data. *Journal of International Business Studies*, 46(6): 724–731.

Jaffe, A. B., & de Rassenfosse, G. 2017. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6): 1360–1374.

Jaffe, A. B., & Lerner, J. 2004. *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It*. Princeton: Princeton University Press.

Kaas, L., & Manger, C. 2012. Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review*, 13(1): 1–20.

Konara, P., & Mohr, A. 2019. Why we should stop using the Kogut and Singh Index. *Management International Review*, 59(3): 335–354.

Kotabe, M. 1992. A comparative study of US and Japanese patent systems. *Journal of International Business Studies*, 23(1): 147–168.

Lehmann-Hasemeyer, S., & Streb, J. 2018. Discrimination against foreigners. The Wuerttemberg patent law in administrative Practice. *Priority Programme 1859 Working Paper Series* No 7.

Lemley, M. A., & Moore, K. A. 2004. Ending abuse of patent continuations. *Boston University Law Review*, 84(1): 63–124.

Lemley, M. A., & Shapiro, C. 2005. Probabilistic patents. *Journal of Economic Perspectives*, 19(2): 75–98.

LeVine, R. A., & Campbell, D. T. 1972. *Ethnocentrism*. New York: John Wiley.

Liang, M., 2012. Chinese patent quality: Running the numbers and possible remedies. John Marshall Review of Intellectual Property Law, 11: 478–522.

Liegsalz, J., & Wagner, S., 2013. Patent examination at the State IP office in China. *Research Policy*, 42(2): 552–563.

Lyman, S. M. 2000. The “Yellow Peril” mystique: origins and vicissitudes of a racist discourse. *International Journal of Politics, Culture, and Society*, 13(4): 683–747.

Marco, A. C., Sarnoff, J. D., & Charles, A. W. 2019. Patent claims and patent scope. *Research Policy*, 48(9): 103790.

Martínez, C. 2011. Patent families: When do different definitions really matter? *Scientometrics*, 86(1): 39–63.

Maskus, K. E., & Penubarti, M. 1995. How trade-related are intellectual property rights? *Journal of International Economics*, 39(3–4): 227–248.

Morgan, R., Lundine, J., Irwin, B., & Grépin, K. A. 2019. Gendered geography: an analysis of authors in The Lancet Global Health. *The Lancet Global Health*, 7(12): e1619-e1620.

Ogan, C., Willnat, L., Pennington, R., & Bashir, M. 2014. The rise of anti-Muslim prejudice: Media and islamophobia in Europe and the United States. *International Communication Gazette*, 76(1): 27–46.

Palangkaraya, A., Jensen, P. H., & Webster, E. 2017. The effect of patents on trade. *Journal of International Economics*, 105: 1–9.

Popp, D., Juhl, T., & Johnson, D., 2003. Time in purgatory: Determinants of the grant lag for US patent applications. *Topics in Economic Analysis and Policy*, 4: 1–43.

Prud'homme, D., & Zhang, T. 2019. *China’s Intellectual Property Regime for Innovation*. Switzerland: Springer Nature Switzerland AG, 237 p.

Sampat, B. N., & Amin, T. 2013. How do public health safeguards in Indian patent law affect pharmaceutical patenting in practice? *Journal of Health Politics, Policy and Law*, 38(4): 735–755.

Sumner, W. G. 1906. *Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals*. New York: Ginn and Company, 692 pages.

Tong, T., Zhang, K., He, Z. L., & Zhang, Y. C., 2018. What determines the duration of patent examination in China? An outcome-specific duration analysis of invention patent applications at SIPO. *Research Policy*, 47(3): 583–591.

Webster, E., Jensen, P. H., & Palangkaraya, A. 2014. Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics*, 45(2): 449–469.

Yang, D. 2008. Pendency and grant ratios of invention patents: A comparative study of the US and China. *Research Policy*, 37(6–7): 1035–1046.

Yang, D., 2019. National treatment, institutions, and IP uncertainties: An analytics of compliance, change and comparability. *International Business Review*, 28(5): 101585.

Yang, D., Sonmez, M., 2018. Global norm of national treatment for patent uncertainties: A longitudinal comparison between the US and China. *Journal of World Business*, 53(2): 164–176.

Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S. 2017. Nationality classification using name embeddings. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1897–1906.

Ye, J., & Skiena, S. 2019. The Secret Lives of Names? Name Embeddings from Social Media. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3000–3008.

APPENDIX A. DATASET DESCRIPTION

This appendix provides a brief overview of the dataset used for the analysis as well as all the intermediary tables that we assembled in order to build it.

Data tables

01_ applnId_attorney

This table contains the attorney information for each application ID. The data are extracted from the 'Correspondence_address' table in the USPTO PAIR database, field 'correspondence_name_line_1.' We then did some pre-processing on the names in order to harmonize similar names and to generate a unique identifier. We first removed special characters, including commas and dots. We then applied standardization rules for company names by the Patent Data Project in order to map the different representations of a company's names into a unique representation.[1]

02_appltId_publn

This table contains publication information and number of applicants and inventors for each application ID. The data come from tables 'TLS227_PERS_PUBLN' and 'TLS211_PAT_PUBLN' from the PATSTAT database and correspond to the earliest publication document associated with each application.

03_ appln_inventors_uspto

This table contains the name of inventors, their sequence numbers (*i.e.*, the order of appearance in the patent document), and their countries of residence from the 'application_data' table in the USPTO OCE PAIR database. We use the table for extracting (first name, last name, country code) triples which we feed to NamSor and NamePrism APIs.

04_examiners_uspto

This table contains information on examiners (including name and art unit). The data are extracted from the 'application_data' table in the USPTO OCE PAIR database.

05_ applnInfo_uspto

This table contains information on applications at the USPTO, including the type of application, key dates, examiners, number of office actions and other relevant information. The data are extracted from tables 'application_data' and 'transaction' in the USPTO OCE PAIR database as well as the 'applications' table in USPTO OCE PEDS database.

06_family_customDef

This table contains family information for each application. It is built using priority, technical relation, and continuation data from the PATSTAT database. Appendix B explains the algorithm we have implemented to assign a family ID to each application ID. This special type of family ID assigns the same ID to applications that can be considered as twin inventions.

07_twin_appln

This table contains all the pairs of twins and associated information. It is generated using table '06_family_customDef' by extracting any pairs of (appln_id_1, appln_id_2) that have the same family ID and for

[1] Source: https://sites.google.com/site/patentdataproject/Home/posts/namestandardizationroutinesuploaded

which the patent authority of the first application (appln_id_1) is the USPTO. Additional information is added using table ‘TLS201_APPLN’ from the PATSTAT database.

08_matching_applnNrOrig_applnId

This table matches application numbers to their correspondence application ID. We generate these data using a combination of data from GOOGLE PATENTS, USPTO OCE PAIR and PATSTAT.

09_matching_applnNrPAIR_applnNrOrig

This table matches the special application number found in the PAIR database to the application number printed in the publications issued by the USPTO. It is imported directly from USPTO OCE PAIR.

10_assignee_appln

Contains assignee information per application ID. This table only includes assignees that are not listed as inventors.

11_attorneyPFE_on_assignee

This table contains the attorney pseudo fixed effect information for each assignee ID. We used data on attorneys extracted from USPTO PAIR and assignee information from PATSTAT. We calculated the attorney pseudo fixed effect for each assignee by computing the average grant rate of all applications represented by that specific attorney for all other assignees (*i.e.*, excluding the applications of the current assignee).

12_artunitPFE_on_assignee

This table contains the art unit pseudo fixed effect information for each assignee. We used data on art unit extracted from USPTO PAIR and assignee information from PATSTAT. We then calculated the art unit pseudo fixed effect for each assignee by computing the average grant rate of all applications processed in that specific art unit for all other assignees (*i.e.*, excluding the applications of the current assignee).

13_examinerPFE_on_assignee

This table contains the examiners pseudo fixed effect information for each assignee. We used data on examiners extracted from USPTO PAIR and assignee information from PATSTAT. We then calculated the examiner pseudo fixed effect for each assignee by computing the average grant rate of all applications reviewed by that specific examiner for all other assignees (excluding the applications of the current assignee).

14_name_ethnicity

This table contains the predicted ethnicity information for each pair of (first name, last name). We first extracted all the unique (first name, last name) pairs for examiners and inventors. We have then used the ‘NamePrism API’ to predict the most likely ethnicity for each pair. NamePrism uses name embeddings, which is then used to classify names into different nationalities and ethnicities. More information about this API can be found at the following URL: http://www.name-prism.com/api.

15_name_genders

This table contains the predicted gender information for each triple of (first name, last name, country code). We first extracted all the unique (first name, last name, country code) triples from examiner and inventor tables. We have then used the ‘NamSor API v2’ to predict the most likely gender for each triple. When the country code information was not available, we have used only (first name, last name) pairs for the prediction. NamSor software relies on sociolinguistics and machine learning models to classify names by gender, origin, and ethnicity. More information about this API can be found at the following URL: https://www.namsor.com/.

16_name_origin

This table contains the predicted country of origin information for each triple of (first name, last name, country code). The construction procedure is similar to that adopted for creating table '15_name_gender.'

17_appln_examInvtOrigin

This table contains the various indicators about the country of origin and country of residence of the examiners and inventors. The backbone of this table is extracted from USPTO PAIR and the indicators are extracted mainly from the tables constructed using NamSor and NamePrism APIs.

18_appln_portfolioSize

This table contains information about the size of the patent portfolio by the applicant for each application. The portfolio size is the number of applications that has been filed by the applicant within the past five years at the time of filing the focal application.

19_appln_grantInfo

This table contains the information regarding the grant outcome of applications in different jurisdictions or different families, sourced from PATSTAT.

20_appln_pct

This table contains the information regarding whether the application has been filed through PCT route or not. We have sourced the initial data from PATSTAT.

21_appln_claims_google

This table contains the information about the number of independent claims and the average number of words in each of the independent claims. The numbers are extracted using a python script that has been calculated for the first publication of each application. We sourced the initial data from GOOGLE PATENTS.

22_examInvt_cultDist

This table contains the information regarding the minimum cultural distance between examiners and inventors. It is based on the work of Hofstede (1980) introducing 4 cultural dimensions of different countries around the world. We have used the updated version of 2013 with 6 cultural dimensions (link). The final table is then constructed using the formula proposed by Konara and Mohr (2019) as:

$$CD_{i,j} = \sqrt{\sum_{k=1}^{6} \left(I_{ki} - I_{kj}\right)^2 / 6\, V_k}$$

$CD_{i,j}$: $cultural\ distance\ between\ two\ countries\ i\ and\ j\ based\ on\ the\ standardized\ euclidean\ distance$

I_{ki} : $index\ for\ the\ k^{th}\ Hofstede\ cultural\ dimension\ and\ i^{th}\ country$

V_k : $variance\ of\ index\ k^{th}\ dimension$

23_final_table

This table aggregates all the information available in other tables for the final econometric analysis.

Description of attributes

Attribute name	Data type	Table(s) containing the attribute	Original source	Description
abandon_date	DATE	[''23_final_table', '05_applnInfo_uspto']	USPTO PAIR	The date at which the application has been abandoned. ‘0’ indicates that it has not been abandoned (yet) and ‘NA’ indicates that there is no information available.
appln_auth	STRING	['06_family_customDef']	PATSTAT	Application authority responsible for processing the patent application
appln_auth_1	STRING	['07_twin_appln']	PATSTAT	Application authority responsible for processing the patent application of the first twin in the exact twin pairs
appln_auth_2	STRING	['07_twin_appln']	PATSTAT	Application authority responsible for processing the patent application of the second twin in the exact twin pairs
appln_filing_year_1	INT64	['07_twin_appln']	PATSTAT	Year in which the first application of the twins’ pair has been filed
appln_filing_year_2	INT64	['07_twin_appln']	PATSTAT	Year in which the second application of the twins’ pair has been filed
appln_filing_year_US	INT64	[' '19_appln_grantInfo', '23_final_table']	PATSTAT	Year in which the U.S. application of the twins’ pair has been filed
appln_id	INT64	['08_matching_applnNrOrig_applnId', '03_ appln_inventors_uspto', '17_appln_examInvtOrigin', '18_appln_portfolioSize', '01_ applnId_attorney', '02_ applnId_publn', '10_ pureAssignee_appln', '03_ appln_inventors_uspto_update', '05_applnInfo_uspto', '06_family_customDef']	PATSTAT	Unique application ID as defined in PATSTAT
appln_id_1	INT64	['07_twin_appln']	PATSTAT	Same as ‘appln_id’, but only created for the first application for each twin
appln_id_2	INT64	['07_twin_appln']	PATSTAT	Same as ‘appln_id’, but only created for the second application for each twin
appln_id_US	INT64	[' '19_appln_grantInfo', '23_final_table']	PATSTAT	Please refer to ‘appln_id’

appln_nr_PAIR	STRING	['09_matching_applnNrPAIR_applnNrOrig']	USPTO PAIR	Application number defined in the USPTO PAIR dataset for each USPTO application
appln_nr_orig	STRING	['09_matching_applnNrPAIR_applnNrOrig', '08_matching_applnNrOrig_applnId']	PATSTAT	Application number that has been issued by the patent authority where the national, international or regional application was filed.
appln_type	STRING	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Identifies the type of an application as either a regular nonprovisional, provisional, re-issue, re-examination, or PCT application
applt_seq_nr	INT64	['10_ pureAssignee_appln']	PATSTAT	Number indicating the place in the list of applicants in the application
artunit_FE	FLOAT64	[' '23_final_table']	USPTO PAIR	Pseudo fixed effect calculated on the Art Unit variable
assignee_han_name	STRING	['10_ pureAssignee_appln']	PATSTAT	This field contains for many applicants the names as harmonized by the OECD HAN (Harmonized Applicant Name) project of the OECD (For more information please refer to the PATSTAT data catalogue)
assignee_id	INT64	['12_ artunitFE_on_assignee', '11_ attorneyFE_on_assignee', '13_examinerFE_on_assignee', '10_ pureAssignee_appln', '23_final_table']	USPTO PAIR	The surrogate key based on the elements in the alternate primary key of table 'TLS206_PERSON' of PATSTAT for the assignees (For more information, please refer to the PATSTAT data catalogue)
assignee_sector	STRING	['10_ pureAssignee_appln']	USPTO PAIR	Same field as 'person_id' in PATSTAT, but only for 'persons' that are pure applicants (meaning that are not also listed as inventors)
attorney	STRING	['01_ applnId_attorney']	USPTO PAIR	The harmonized attorney names (extracted from 'correspondence_address' table in USPTO PAIR and then harmonized)
attorney_FE	FLOAT64	["23_final_table']	USPTO PAIR / PATSTAT	Attorney pseudo fixed effect on each assignee extracted by using attorney data from USPTO PAIR and grant outcome from PATSTAT.
attorney_country_code	STRING	['01_ applnId_attorney']	USPTO PAIR	Same as 'correspondence_country_code' column in 'correspondence_address' table (USPTO PAIR)

attorney_id	STRING	['11_ attorneyFE_on_assignee', '01_ applnId_attorney']	USPTO PAIR	Unique attorney ID created from harmonized attorney names, which were extracted from USPTO PAIR.
attorney_id_US	STRING	[' '23_final_table']	USPTO PAIR	Please refer to ‘attorney_id’
attorney_region_code	STRING	['01_ applnId_attorney']	USPTO PAIR	Same as ‘correspondence_region_code’ which can be found in ‘correspondence_address’ table in USPTO PAIR.
avg_claimWords	FLOAT	['21_appln_claims_google', '23_final_table']	GOOGLE PATENTS	Average number of words per independent claims
country_code	STRING	['15_name_gender', '16_name_origin']	USPTO PAIR	Country code of the names, which were extracted from ‘inventor_country_code’ column of ‘all_inventors’ table in USPTO PAIR
country_origin	STRING	['16_name_origin']	NamSor API	Predicted country of origin for each triple of (first name, last name, country code) using NamSor API
disposal_type	STRING	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR (v2015)	Disposal type, can take values ‘issued’, ‘pending’, or ‘abandoned’
docdb_avg_grant	FLOAT64	[' '19_appln_grantInfo', '23_final_table']	PATSTAT	Average grant rate calculated over DOCDB families for 8 application authorities including 'EP', 'JP', 'CN', 'KR', 'DE', 'CA', 'AU', and 'TW'.
ethnicity	STRING	['14_name_ethnicity']	NamePrism API	Predicted ethnicity for each pair of (first_name, last_name) using NamePrism API
examiner_FE	FLOAT64	[' '23_final_table']	USPTO PAIR	Examiner pseudo fixed effect on each assignee extracted by using examiner data from USPTO PAIR and grant outcome from PATSTAT.
examiner_art_unit	STRING	[''04_examiners_uspto', '12_ artunitFE_on_assignee', '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	The art unit to which the examiner belongs to. It was extracted from ‘application_data’ table in USPTO PAIR.
examiner_chinese	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for examiners who have ‘China’ as their predicted country of origin (‘country_origin’)
examiner_female	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for examiners being female (Extracted from ‘predicted_gender’ variable).

examiner_id	STRING	[' '04_examiners_uspto', '13_examinerFE_on_assignee', '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Unique examiner ID. It is the same as ‘examiner_id’ which can be found in ‘application_data’ table in USPTO PAIR dataset.
examiner_muslim	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamePrism API / USPTO PAIR	Flag for examiners being Muslim based on the predicted ethnicity from the (first_name, last_name) pairs (extracted from USPTO PAIR) using NamePrism API
examiner_name_first	STRING	['04_examiners_uspto']	USPTO PAIR	Examiner’s first name extracted from ‘application_data’ table.
examiner_name_last	STRING	['04_examiners_uspto']	USPTO PAIR	Examiner’s last name extracted from ‘application_data’ table.
examiner_name_middle	STRING	['04_examiners_uspto']	USPTO PAIR	Examiner’s middle name extracted from ‘application_data’ table.
family_id	INT64	[' '19_appln_grantInfo', '23_final_table', '07_twin_appln', '06_family_customDef']	PATSTAT	Custom family ID, constructed using our algorithm, which places the exact twins in the same family ID. For more information about the algorithm, please refer to Appendix B.
filing_date	DATE	[' '23_final_table', '05_applnInfo_uspto']	PATSTAT	The date t which the application has been filed (same as ‘appln_filling_date’ in PATSTAT)
gender_scale	FLOAT64	['15_name_gender']	NamSor API	Scale of predicted gender for each (first name, last name, country code) triple.
grant_date	DATE	[' '23_final_table', '05_applnInfo_uspto']	USPTO PEDS	Grant date extracted from ‘applications’ table in USPTO PEDS.
grant_rate	FLOAT64	['12_ artunitFE_on_assignee', '11_ attorneyFE_on_assignee', '13_examinerFE_on_assignee']	PATSTAT and USPTO PAIR	Average grant rate that indicates the calculated pseudo fixed effect on each assignee. For each of ‘examiner_FE’ (examiner pseudo fixed effect), ‘attorney_FE’, and ‘artunit_FE’ please refer to the corresponding variables.
granted_1	BOOL	['07_twin_appln']	PATSTAT	Grant outcome of the first application in each twin’ pair (Extracted from ‘granted’ column in ‘TLS201_APPLN’ table in PATSTAT).
granted_2	BOOL	['07_twin_appln']	PATSTAT	Grant outcome of the second application in each twin’s pair (Extracted from ‘granted’ column in ‘TLS201_APPLN’ table in PATSTAT).

granted_US	INT64	[' '19_appln_grantInfo', '23_final_table']	PATSTAT	Grant outcome of the US application (Extracted from ‘granted’ column in ‘TLS201_APPLN’ table in PATSTAT).
invt_angSax	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	USPTO PAIR / Other Sources	Flag for inventors living in Anglo-Saxon countries, including ‘US’, ‘UK’, ‘CA’, ‘AU’, and ‘NZ’
invt_chinese	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for inventors who have ‘China’ as their predicted country of origin (‘country_origin’)
invt_country_code	STRING	['03_appln_inventors_uspto', '03_appln_inventors_uspto_update']	USPTO PAIR	Please refer to ‘country_code’.
invt_country_origin	STRING	['03_appln_inventors_uspto_update']	NamSor API	Please refer to ‘country_origin’
invt_eastasian	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamePrism API	Flag for inventors who have ‘EastAsian’ as their ethnicity (Please refer to ‘ethnicity’).
invt_eng	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	USPTO PAIR / Other Sources	Flag for inventors who reside in an English speaking country (including countries with English as their official/educational language). Reference: https://en.wikipedia.org/wiki/List_of_territorial_entities_where_English_is_an_official_language)
invt_ethnicity	STRING	['03_appln_inventors_uspto_update']	NamePrism API	Please refer to ‘ethnicity’
invt_female	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for female inventors (Extracted from ‘invt_gender’ variable).
invt_foreign	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	USPTO PAIR	Flag for inventors who are not living in the United States. (Extracted from ‘invt_country_code’)
invt_gdppc	FLOAT64	[' '17_appln_examInvtOrigin', '23_final_table', '03_appln_inventors_uspto_update']	USPTO PAIR / Other Sources	GDP per capita of the country of residence of inventor (Source: https://data.worldbank.org/indicator/ny.gdp.pcap.pp.cd)
invt_gender	STRING	['03_appln_inventors_uspto_update']	NamSor API	Please refer to ‘predicted_gender’

invt_muslim	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	NamePrism API / USPTO PAIR	Flag for inventors being Muslim based on the predicted ethnicity from the (first_name, last_name) pairs (extracted from USPTO PAIR) using NamePrism API
invt_name_first	STRING	['03_ appln_inventors_uspto', '03_ appln_inventors_uspto_update']	USPTO PAIR	Inventors' first name extracted from 'all_inventors' table.
invt_name_last	STRING	['03_ appln_inventors_uspto', '03_ appln_inventors_uspto_update']	USPTO PAIR	Inventors' last name extracted from 'all_inventors' table.
invt_name_middle	STRING	['03_ appln_inventors_uspto', '03_ appln_inventors_uspto_update']	USPTO PAIR	Inventors' middle name extracted from 'all_inventors' table.
invt_res_china	INT64	[' '17_appln_examInvtOrigin', '23_final_table']	USPTO PAIR	Flag for inventors who reside in China (extracted using 'country_code')
invt_seq_nr	STRING	['03_ appln_inventors_uspto', '03_ appln_inventors_uspto_update']	USPTO PAIR	Same as 'inventor_rank' in 'all_inventors' table in USPTO PAIR.
invt_sub_reg_origin	STRING	['03_ appln_inventors_uspto_update']	NamSor API / USPTO PAIR	Inventor's sub region origin predicted using (first_name, last_name) pairs (extracted from USPTO PAIR) using NamSor API.
is_pct	INT64	['20_appln_pct', '23_final_table']	PATSTAT	Flag for applications which has been filed through PCT route
is_US_resident	INT64	['03_ appln_inventors_uspto', '03_ appln_inventors_uspto_update']	USPTO PAIR	Flag for inventors who reside in the United Staetst (Extracted from 'person_country_code' in 'TLS906_PERSON' table)
issue_date	DATE	['23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Patent issue date extracted from 'patent_issue_date' in 'application_data' table.
min_culturalDist	FLOAT	['23_final_table', '22_examInvt_cultDist']	USPTO / Other Sources	Minimum cultural distance between examiner and inventors per application. The initial cultural dimensions were extracted from (Hofstede, 1980) https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

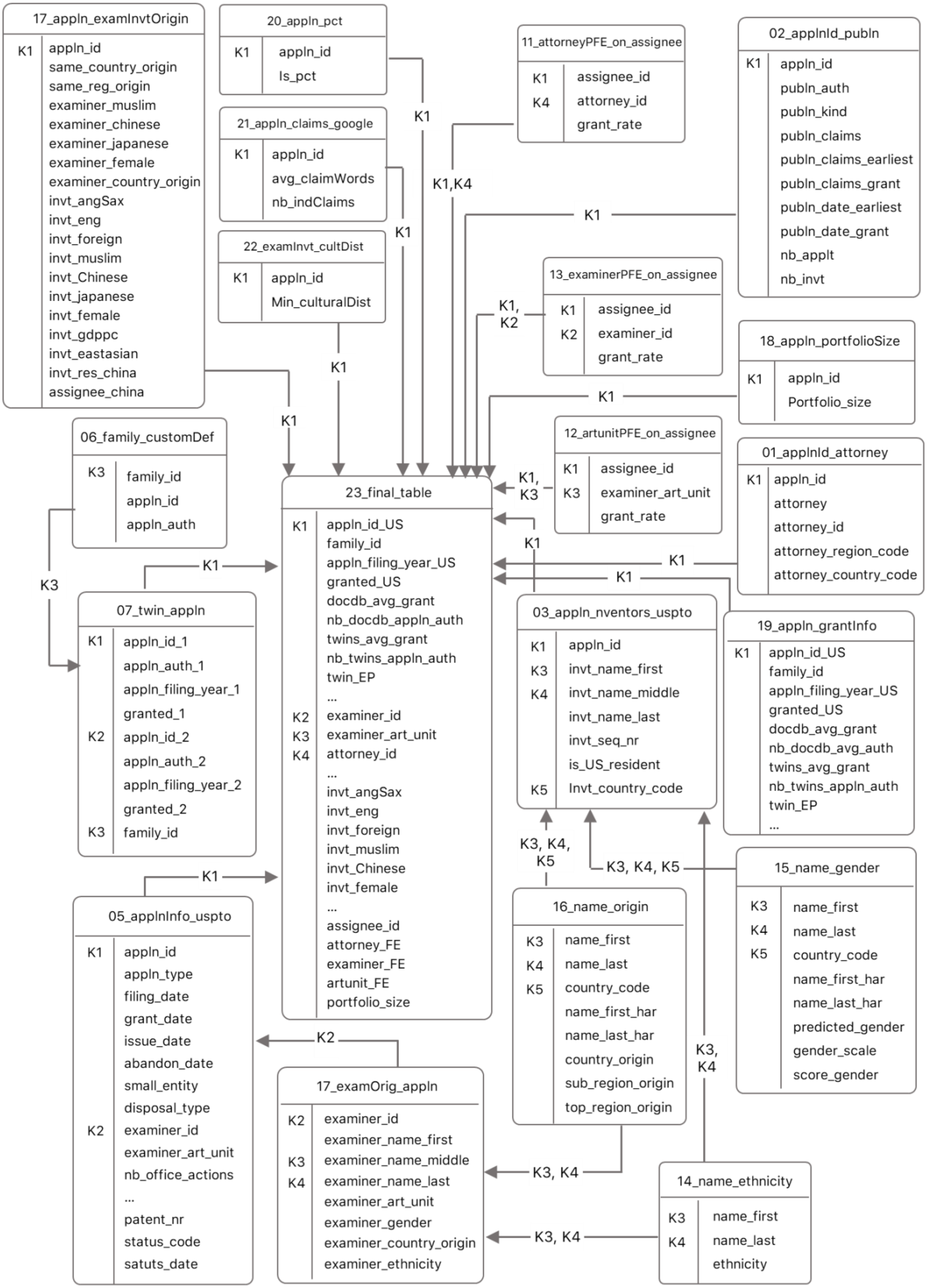
name_first	STRING	['15_name_gender', '14_name_ethnicity', '16_name_origin']	USPTO PAIR	Person first name extracted from ‘application_data’ table.
name_first_har	STRING	['15_name_gender', '16_name_origin']	USPTO PAIR	Person harmonized first name extracted from ‘name_first’ to be used for NamSor and NamePrism APIs.
name_last	STRING	['15_name_gender', '14_name_ethnicity', '16_name_origin',]	USPTO PAIR	Person last name extracted from ‘application_data’ table.
name_last_har	STRING	['15_name_gender', '16_name_origin']	USPTO PAIR	Person harmonized last name extracted from ‘name_last’ to be used for NamSor and NamePrism APIs.
nb_applt	INT64	['02_ applnId_publn']	PATSTAT	Number of applicants constructed using ‘applt_seq_nr’ (using ‘TLS227_PERS_PUBLN’ table) for each application.
nb_applt_US	INT64	['23_final_table']	PATSTAT	Number of applicants for applications in USPTO (Please refer to ‘nb_applt’)
nb_docdb_appln_auth	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	Number of distinct authorities found in each DOCDB family.
nb_indClaims	INT64	['21_appln_claims_google', '23_final_table']	GOOGLE PATENTS	Number of independent claims of the first publication
nb_invt	INT64	[' '02_ applnId_publn']	PATSTAT	Number of inventors constructed using ‘invt_seq_nr’ (using ‘TLS227_PERS_PUBLN’ table)
nb_invt_US	INT64	[' '23_final_table']	PATSTAT	Please refer to ‘nb_invt’
nb_office_actions	INT64	[' '23_final_table', '05_applnInfo_uspto']	USPTO PEDS	Number of office actions per application constructed using ‘applications’ table.
nb_rejection	INT64	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Number of non-final and final rejections, constructed using ‘transactions’ table.
nb_transaction	INT64	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Number of all transactions constructed using ‘transactions’ table.
nb_transaction_aa	INT64	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR / Other Sources	Number of all actions initiated by applicants (using categories of most 100 frequent codes in ‘Appendix B: Description of the Transaction History Tab Release’ of USPTO PAIR dataset – Link: https://www.uspto.gov/sites/default/files/documents/Appendix%20B.pdf)

nb_transaction_ex	INT64	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR / Other Sources	Number of all actions initiated by examiners (using categories of most 100 frequent codes in ‘Appendix B: Description of the Transaction History Tab Release’ of USPTO PAIR dataset – Link: https://www.uspto.gov/sites/default/files/documents/Appendix%20B.pdf)
nb_twins_appln_auth	INT64	[' '19_appln_grantInfo', '23_final_table']	PATSTAT	Number of distinct authorities found in each ‘family_id’ created for twins.
patent_nr	STRING	[' '23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Patent number which can be found in ‘application_data’ table.
portfolio_size	INT64	[' '18_appln_portfolioSize', '23_final_table']	PATSTAT	Number of applications which were filed by the same assignee within the last 5 years.
predicted_gender	STRING	['15_name_gender']	NamSor API	The predicted gender based on (first name, last name, country code) triples.
publn_auth	STRING	['02_ applnId_publn']	PATSTAT	Patent Authority that issued the publication of the application (same as ‘publn_auth’ in PATSTAT)
publn_claims_earliest	INT64	['02_ applnId_publn']	PATSTAT	Number of claims of the earliest publication for each application ID
publn_claims_earliest_US	INT64	['23_final_table']	PATSTAT	Please refer to ‘publn_claims_earliest’
publn_claims_grant	INT64	['02_ applnId_publn', '23_final_table']	PATSTAT	Number of claims of the publication at the time of first grant (constructed using ‘publn_first_grant’ flag in ‘TLS211_PAT_PUBLN’ table)
publn_date_earliest	DATE	['02_ applnId_publn']	PATSTAT	Date of earliest publication for each application ID.
publn_date_earliest_US	DATE	['23_final_table']	PATSTAT	Please refer to ‘publn_date_grant’
publn_date_grant	DATE	['02_ applnId_publn']	PATSTAT	Date of publication at the time of first grant
publn_date_grant_US	DATE	['23_final_table']	PATSTAT	Please refer to ‘publn_date_grant’
publn_kind	STRING	['02_ applnId_publn']	PATSTAT	Publication kind attributed by the Patent Authority issuing the publication (Same as ‘publn_kind’ in ‘TLS211_PAT_PUBLN’ table.
same_country_origin	INT64	['17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for whether the examiner and the strict majority of inventors have the same

				country of origin (constructed using ‘country_origin’)
same_reg_origin	INT64	['17_appln_examInvtOrigin', '23_final_table']	NamSor API	Flag for whether the examiner and the strict majority of inventors have the same region of origin (using ‘sub_region_origin’).
score_gender	FLOAT64	['15_name_gender']	NamSor API	Score of predicted gender returned by NamSor API.
small_entity	STRING	['23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Indicator of whether applicants are considered a small entity. Applications from individual inventors, non-profit firms, and for-profit firms with fewer than 500 employees are granted small-entity status (Same as ‘small_entity_indicator’ in ‘application_data’ table).
status_code	STRING	['23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Same as ‘status_code’ in ‘transactions’ table.
status_date	DATE	['23_final_table', '05_applnInfo_uspto']	USPTO PAIR	Same as ‘status_date’ in ‘transactions’ table.
sub_region_origin	STRING	['16_name_origin']	NamSor API	Predicted sub-region from NamSor API that each (first_name, last_name) pair belongs to.
top_region_origin	STRING	['16_name_origin']	NamSor API	Predicted top-region from NamSor API that each (first_name, last_name) pair belongs to.
twin_AU	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in Australia. Number ‘1’ means that for the corresponding US application, there is a twin application (with the same ‘family_id’) in ‘AU’ jurisdiction which has been granted the patent. ‘0’ means that there is a twin but has not been granted any patents. And ‘-1’ means that there is an application in ‘AU’ jurisdiction within the DOCDB family, corresponding to the US application.
twin_CA	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in Canada. Please refer to ‘twin_AU’ for more information

twin_CN	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in China. Please refer to ‘twin_AU’ for more information
twin_DE	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in Germany. Please refer to ‘twin_AU’ for more information
twin_EP	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins at the EPO. Please refer to ‘twin_AU’ for more information
twin_JP	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in Japan. Please refer to ‘twin_AU’ for more information
twin_KR	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in South Korea.. Please refer to ‘twin_AU’ for more information
twin_TW	INT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	An indicator of the grant outcome of the US twins in Taiwan. Please refer to ‘twin_AU’ for more information
twins_avg_grant	FLOAT64	['19_appln_grantInfo', '23_final_table']	PATSTAT	The average grant rate calculated over ‘family_id’ families for eight application authorities including 'EP', 'JP', 'CN', 'KR', 'DE', 'CA', 'AU', and 'TW'.

Relational schema


17_appln_examInvtOrigin
K1
appln_id
same_country_origin
same_reg_origin
examiner_muslim
examiner_chinese
examiner_japanese
examiner_female
examiner_country_origin
invt_angSax
invt_eng
invt_foreign
invt_muslim
invt_Chinese
invt_japanese
invt_female
invt_gdppc
invt_eastasian
invt_res_china
assignee_china
20_appln_pct
K1
appln_id
Is_pct
21_appln_claims_google
K1
appln_id
avg_claimWords
nb_indClaims
22_examInvt_cultDist
K1
appln_id
Min_culturalDist
11_attorneyPFE_on_assignee
K1
assignee_id
K4
attorney_id
grant_rate
02_applnId_publn
K1
appln_id
publn_auth
publn_kind
publn_claims
publn_claims_earliest
publn_claims_grant
publn_date_earliest
publn_date_grant
nb_applt
nb_invt
13_examinerPFE_on_assignee
K1
assignee_id
K2
examiner_id
grant_rate
18_appln_portfolioSize
K1
appln_id
Portfolio_size
12_artunitPFE_on_assignee
K1
assignee_id
K3
examiner_art_unit
grant_rate
01_applnId_attorney
K1
appln_id
attorney
attorney_id
attorney_region_code
attorney_country_code
06_family_customDef
K3
family_id
appln_id
appln_auth
23_final_table
K1
appln_id_US
family_id
appln_filing_year_US
granted_US
docdb_avg_grant
nb_docdb_appln_auth
twins_avg_grant
nb_twins_appln_auth
twin_EP
...
K2
examiner_id
K3
examiner_art_unit
K4
attorney_id
...
invt_angSax
invt_eng
invt_foreign
invt_muslim
invt_Chinese
invt_female
...
assignee_id
attorney_FE
examiner_FE
artunit_FE
portfolio_size
07_twin_appln
K1
appln_id_1
appln_auth_1
appln_filing_year_1
granted_1
K2
appln_id_2
appln_auth_2
appln_filing_year_2
granted_2
K3
family_id
03_appln_nventors_uspto
K1
appln_id
K3
invt_name_first
K4
invt_name_middle
invt_name_last
invt_seq_nr
is_US_resident
K5
Invt_country_code
19_appln_grantInfo
K1
appln_id_US
family_id
appln_filing_year_US
granted_US
docdb_avg_grant
nb_docdb_avg_auth
twins_avg_grant
nb_twins_appln_auth
twin_EP
...
05_applnInfo_uspto
K1
appln_id
appln_type
filing_date
grant_date
issue_date
abandon_date
small_entity
disposal_type
K2
examiner_id
examiner_art_unit
nb_office_actions
...
patent_nr
status_code
satuts_date
16_name_origin
K3
name_first
K4
name_last
K5
country_code
name_first_har
name_last_har
country_origin
sub_region_origin
top_region_origin
15_name_gender
K3
name_first
K4
name_last
K5
country_code
name_first_har
name_last_har
predicted_gender
gender_scale
score_gender
17_examOrig_appln
K2
examiner_id
examiner_name_first
K3
examiner_name_middle
K4
examiner_name_last
examiner_art_unit
examiner_gender
examiner_country_origin
examiner_ethnicity
14_name_ethnicity
K3
name_first
K4
name_last
ethnicity
K1
K1,K4
K1
K1
K1,
K2
K1
K1
K1
K1
K1,
K3
K1
K1
K1
K3
K1
K1
K3, K4,
K5
K3, K4, K5
K3,
K4
K2
K3, K4
K3, K4

APPENDIX B: ALGORITHM FOR CONSTRUCTING TWIN FAMILIES

We treat the problem of identifying a patent family as a special case of a Set Union problem. To this aim, we need two pieces of information. First, the set of all available applications to be considered (the ‘universe’) and, second, information about the relationship between the nodes (or patent linkages).

We can consider the problem of finding the patent families on a directed acyclic graph $G = (V, E)$, as finding (weakly) connected components (*i.e.*, patent families in this case) using the list of all edges E (*i.e.*, patent linkages) for all nodes N (*i.e.*, all applications). With this formulation in mind, let us first consider the Pseudo-code for the *Union Find* algorithm:

ALGORITHM 1

```
1. Initialize an array A with the nodes N at its IDs and also its entries (the ID and the entry are
   initially the node’s value). This step is the same as initializing the singleton sets, in which each
   node is pointing to itself $p(n_i) = n_i$.
2. For each edge $e = (n_i, n_j)$ in the edge list E:
    a. Find all IDs that are pointing to the same number as $p(n_i)$ and change them to where
       node $n_j$ is pointing, i.e. $p(n_j)$.
3. Return the list of IDs and their associated pointers as the connected components.
```

We will not analyze the running time of this algorithm, but by considering ‘n’ nodes and ‘m’ edges, the running time of this algorithm will then be in $O(n^2)$.[1] Bearing in mind that in the worst case the number of union operations will be equal to $m = n - 1$, which is the maximum number of edges in an acyclic graph.

We first consider the case of constructing the INPADOC family using the list of all applications from PATSTAT (table ‘*TLS201_APPLN’*) as the list of nodes in our problem. Concerning the list of relations, we consider the patent linkage from the following three PATSTAT tables: Paris Convention priorities (table ‘*TLS204_APPLN_PRIOR*’); Technical (table ‘*TLS205_TECH_REL*’); Application Continuation (table ‘*TLS216_APPLN_CONTN*’).

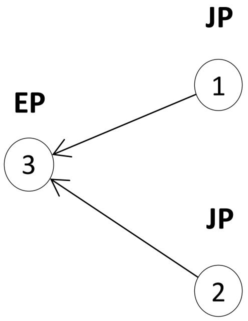
Implementing ALGORITHM 1 with a few tweaks and using the three mentioned linkages tables, results into the INPADOC extended patent families (Martinez 2010).[2]

However, the members of INPADOC family cannot be treated as the ‘same invention’ (*i.e.*, exact twins). There are many reasons for this. For example, applicants filing for international patent protection through the PCT are able to amend patent claims in each jurisdiction. Therefore, it is possible that the claimed inventions vary across different patent offices. However, for the purpose of this paper, we address other issues. To illustrate the problem, we will use two imaginary examples. After introducing the algorithm, we will also investigate some real cases. Let us start with the following simple example.

[1] Tarjan, R.E. and Van Leeuwen, J., 1984. Worst-case analysis of set union algorithms. Journal of the ACM (JACM), 31(2): 245–281.

[2] Martinez, C., 2010. Insight into different types of patent families. OECD Science, Technology and Industry

Figure 1. First example

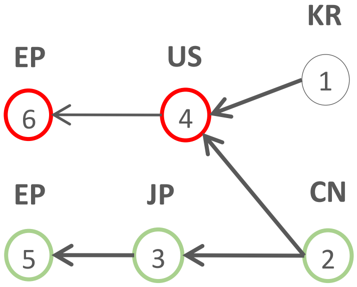


The arrows in Figure 1 are sourcing from the 'priority filings' and pointing towards the 'secondary filings.' The figure illustrates two applications that were filed first in Japan. The applicant then decided to also protect the invention at the European Patent Office by combining the two JP applications into one EP application, which is presumably of larger scope than any of the individual JP applications. It is clear that although these applications are linked by priority filings, none of them covers the same invention—but they are considered as one family under the INPADOC definition.

The twin algorithm will explicitly exclude such cases. More generally, multiple applications that are filed in the same jurisdiction and are linked to one patent (or a set of patents) in another jurisdiction, are considered either a division or a merger of other patents. They are consequently excluded from the set of twins.

To further illustrate the concept of twin patents, let us consider a more complex case. Imagine that an applicant has two priority applications, one in Korea (KR) and one in China (CN). It then decides to protect application '2' also in Japan and to combine applications '1' and '2' to file one patent with a broader scope at the USPTO. Later, the company also decides to protect each of those patents at the EPO. Here, it is clear that applications (2,3,5) can be considered as exact twins and that applications (4,6) can be considered another set of exact twins. But application '1' should not be considered as an exact twin to any other applications, or application '4' should not be considered as the exact twin of patent '3' or '2', since it is a combination of the applications '1' and '2.

Figure 2. Second example



Clearly, applications in one jurisdiction that have the same set of priorities or are the priorities of a same set of applications, must be dealt with extra care. Here, we take the more conservative approach of assigning a different family ID to these applications, although it might be possible that some of these cases are still covering the same inventions (as we will show in a real example later on).

The identification of twins relies on a modified version of ALGORITHM 1. Below, we describe the logic of the algorithm, in SQL mode. The full implementation (with small tweaks for implementing on Big Query) is available in our GitHub repository.[3]

ALGORITHM 2

1. Initialize the 'Relation' table with four columns ('*Parent_ID*', '*Parent_Authority*', '*Prior_Set*', '*Children_Set*') using three patent linkage tables 'TLS204_APPL_PRIOR', 'TLS205_TECH_REL', and 'TLS216_APPLN_CONTN' and the application authority data from 'TLS201_APPLN' table. The '*Parent_ID*' and '*Parent_Authority*' are the same as '*Application_ID*' and '*Application_Authority*', respectively. The '*Prior_Set*' is the set of all priority filings that each parent id is pointing to. The '*Children_Set*' is the set of all '*Parent_ID*' that are from the same '*Parent_Authority*' and have the same '*Prior_Set*'.
   For applications that are in the table 'TLS201_APPLN' but not in the 'Relation' table, add their data to the 'Relation' table by setting '*Prior_Set*' and '*Children_Set*' initially containing only the '*Parent_ID*' as their member.
2. Initialize the 'Family' table with three columns as ('*Application_ID*', '*Application_Authority*', '*Parent_Set*'), where '*Application_ID*' and '*Application_Authority*' are the same as '*appln_id*' and '*appln_auth*' columns from 'TLS201_APPLN', respectively. And initially, '*Parent_Set*' is the set containing only its '*Application_ID*' as its member.
3. While there exists a '*Parent_Set*' in the 'Family' table that is updated:
   a. For each '*Application_ID*', update the *parent ID*s in the '*Parent_Set*' using ('*Parent_ID*', '*Prior_Set*') pairs in 'Relation' table, only if the initial '*Parent Set*' (at the beginning of step 3) is a subset of the '*Children Set*' (for those application IDs that are pointing to several priors, add all of them to the parent set). Flag the parent sets that have been changed.
4. Assign a unique family ID to each distinct '*Parent_Set*' (applications with the same parent set will be located in the same family).
5. Return the final 'Family' table.

Note that the running speed of the algorithm can be improved by first removing all the isolated nodes (patents without any links in the patent linkage tables) and then adding them to the final family table after running the algorithm.

It can be seen that ALGORITHM 2 is similar to ALGORITHM 1, except for two parts. First, the steps that are updating the family IDs has been modified to exclude some applications that were considered previously to belong to the same family. As explained before, this is because we do not want to put all applications from the same authority in one family that are either priority to the same set of applications or that have the same set of priority applications. In addition, we also want to exclude cases where a patent is split into several patents in another jurisdiction for the secondary filing or if several priority applications of the same jurisdiction were combined and filed as one patent in another jurisdiction. We will explore this intuition by using a real example.

Second, in order to optimize for the BigQuery implementation, we changed the union step and added an outer 'While' loop. For the instance of finding patent families, this While loop will not cause a problem. The number of times that it will iterate depends on the 'chain' of priority filings (*i.e.*, the longest shortest path for all the components). By assuming that the longest shortest path in all components is c and assuming $c \ll n$, the effect of While loop can be ignored in analyzing a large number of patents. For example, if the longest chain of priority filings contains 5 applications, where application '1' is the priority of application of '2', application '2' is priority of '3', and so on, the chain of priority filing will be $(1, 2, 3, 4, 5)$ and for this specific case, we need to iterate 5 times. The assumption of having $c \ll n$ is

[3] The GitHub repository can be accessed at the following URL: https://github.com/rezaho/uspto_2019

reasonable in our case, since we are dealing with more than 90 million applications and the chain of priority filings is might not exceed a single digit.

Table 1. Real example from a simple case with only one level of priority filings

Application ID	Application Authority	Family ID	Application Number	Filing Year	Granted	INPADOC Family ID	Prior Application ID
274702168	CN	72673	200880011231	2008	TRUE	39836	55233300 55544732
54949232	EP	72673	8737383	2008	TRUE	39836	55233300 55544732
274357034	US	72673	45019108	2008	TRUE	39836	55233300 55544732
72673	WO	72673	2008000812	2008	FALSE	39836	55233300 55544732
55233300	JP	55233300	2007100080	2007	TRUE	39836	NA
55544732	JP	55544732	2007127128	2007	TRUE	39836	NA

The information in Table 1 is extracted from the result of our algorithm on actual PATSTAT data. Table '*06_family_customDef*' contains the final family information, as mentioned in Appendix A. We can see that all of these applications are recognized as one INPADOC family ID. However, only the first four were recognized as one family in our family definition. We sought to understand the validity of our algorithm by manually looking at the publications of applications and compared their claims.[4]

The result of this comparison can be found in Table 2. It is clear from this table that all of the applications that were recognized as exact twins, have exactly the same claims. However, the additional two applications from JPO that are in the same INPADOC family, do not share the same claims. One possible explanation for this case is that the applicant preferred to file the new patent with a broader scope in the new jurisdiction.

Table 2. Comparing different claims of the applications within the same family

Patent Office	JP (*)	JP (*)	US	CN	EP	WO
Publication Number	JP-4265675-B2	JP-4321623-B2	US-8291697-B2	CN-101652551-B	EP-2145092-B1	WO-2008122866-A2
Application Number	2007100080	2007127128	45019108	200880011231	8737383	2008000812
Granted	No	Yes **	Yes **	Yes **	Yes **	Yes **
Abstract	NA	NA	(Base)	Same	NA	Same
1st Claim	1	-	1	1	1	1
2nd Claim	2	-	2	2	2	2
3rd Claim	-	1	3	3	3	3
4th Claim	-	1	4	4	4	4

[4] We went through a large number of cases manually, but only report selected cases for illustration purpose.

Now, let us consider a more complicated example. Consider the INPADOC family ID '1981'. One can query the members of this family using PATSTAT table '*TLS201_APPLN*'. This is a rather large family with 29 members. We will not analyze all of the family members. We only consider the 12 applications that are filed at the USPTO. These applications were all filed by 'Marinus Pharmaceuticals Inc' and all of them are in the same INPADOC family. By looking at their DOCDB family members, we see that still 9 applications are considered in the same DOCDB family (docdb_family_id='38067988'). However, by looking at their actual publications, we can see that their claims and even their titles are different from each other and thus they do not cover the same invention and should not be considered in the same family. We see that our family definition takes the more conservative approach and assigns a different family ID to each of them.

By taking the more conservative approach, we might exclude some patents that should remain in the same family. As an example, in Table 3, the application ID '38037505', which was filed at the JPO, is excluded from the rest by our family definition. However, considering the claims in their priority filings, there are reasons to believe that it should belong to the same family. The same reasoning applies to the applications that are filed in Germany and in China.

Table 3. A more complex example that illustrates the limits of our approach

Row	Application ID	Application Authority	Family ID	INPADOC family ID	DOCDB family ID	Granted	Prior Application ID
0	50519399	US	902975508	1488396	27791046	Yes	38037505 902975508
1	45187860	TW	902975508	1488396	27791046	Yes	38037505 902975508
2	41322288	MX	902975508	1488396	27791046	Yes	38037505 902975508
3	416266165	KR	902975508	1488396	27791046	Yes	38037505 902975508
4	902975508	JP	902975508	1488396	902975508	No	NA
5	17681908	ES	902975508	1488396	27791046	Yes	38037505 902975508
6	15997183	EP	902975508	1488396	27791046	Yes	38037505 902975508
7	4794575	CA	902975508	1488396	27791046	Yes	38037505 902975508
8	3497761	BR	902975508	1488396	27791046	Yes	38037505 902975508
9	2573818	AU	902975508	1488396	27791046	Yes	38037505 902975508
10	38037505	JP	38037505	1488396	27791046	No	902975508
11	14134328	DE	14134328	1488396	27791046	Yes	38037505 902975508
12	14134327	DE	14134327	1488396	27791046	No	38037505 902975508
13	7013329	CN	7013329	1488396	27791046	Yes	38037505 902975508

14	6945392	CN	6945392	1488396	27791046	Yes	38037505 902975508

The advantage of using the proposed algorithm compared to DOCDB is that DOCDB only considers applications with the same priority as one family, while there might be some other applications that are indirectly connected (by a chain of priority filings) to them and cover the same invention, which DOCDB does not consider. Comparing to INPADOC family, which considers all applications that are directly or indirectly connected to each other as one family, our algorithm excludes the multiple applications in the same jurisdictions that are the priority filings of the same set of applications or has the same set of priorities (*e.g.*, if there are two applications pointing to the same priorities and are filed in the same jurisdictions). In addition, we exclude cases when multiple patents are combined or a patent is split into several applications in their secondary filings in other jurisdictions.

We do not claim that our approach is not prone to failure (in the sense of identifying which applications can be treated as exact twins). This is because, in the proposed method, we do not consider the application claims and their semantic meaning to decide which should be considered as one family. Rather, we only consider the relationship (Priority Filing, Secondary Filing) between each two patents, and follow certain rules to exclude certain members that have a high chance of not covering the same invention.