Disambiguation of Patent Inventors and Assignees Using High-Resolution Geolocation Data

Greg Morrison* Massimo Riccaboni† Fabio Pammolli‡

June 8, 2016

Abstract

Patent data represent a significant source of information on innovation, knowledge spillovers, and the evolution of technology through networks of citations, co-invention and co-assignment. A major obstacle to extracting useful information from this data is the problem of name disambiguation: linking alternate spellings of individuals or institutions to a single identifier to uniquely determine the parties involved in the creation of a technology. In this paper, we describe a new algorithm that uses high-resolution geolocation to disambiguate both inventor and assignees on about 8.5 million patents found in the European Patent Office (EPO), under the Patent Cooperation treaty (PCT), and in the US Patent and Trademark Office (USPTO). We show this disambiguation is consistent with a number of ground-truth benchmarks of both assignees and inventors, significantly outperforming the use of undisambiguated names to identify unique entities. A significant benefit of this work is the high quality assignee disambiguation with worldwide coverage coupled with an inventor disambiguation (that is competitive with other state of the art approaches) in multiple patent offices. To our knowledge this is the broadest and most accurate simultaneous disambiguation and cross-linking of the inventor and assignee names for a significant fraction of patents in these three major patent collections.

1 Introduction

In many contexts, technological progress and innovation is essential to national or regional economic growth and output. One way of measuring innovation is the production of patents, which represent a technological advancement produced by individuals (generally,

*Lab. for the Analysis of Complex Economic Systems, IMT Institute for Advanced Studies, Lucca, Italy and Dept. of Physics, The University of Houston, Houston, TX, USA

†IMT Institute for Advanced Studies, Lucca, Italy and Department of Managerial Economics, Strategy and Innovation, K.U. Leuven, Leuven, Belgium.

‡IMT Institute for Advanced Studies, Lucca, Italy

these are inventors on the patents) working at research institutions (generally, these are the assignees on the patent). The analysis of patent databases has provided techniques for evaluating information spillovers [1, 2], inventor mobility between regions [3, 4], interregional and international collaborations [5, 6, 7, 8, 9], and the emergence of new technologies [10]. Among many others, these studies have provided an in-depth picture of the dynamics of regional and institutional talents and quantify the success of various inter-institutional and inter-regional collaboration, of great use to policy makers.

A major problem in the use of patent data (or any bibliometric database, such as for scholarly publications [11]) is the disambiguation of authors or institutions. There are a wide range of alternate spellings of a person's or institution's name, where, for example, "The National Institutes of Health" and "NIH" may refer to the same institution. Typos and misspellings of names are also common in bibliographic data (e.g. "National Institute of Health," missing an "s" in the second word). The goal of disambiguation is to link all of these alternate spellings of institutional or individual names *without* incorrectly linking similar names referring to distinct entities. This is a difficult task, as there are millions of names to disambiguate (making pairwise comparisons of the full dataset computationally expensive) and an evaluation of how likely two names on patents are to be the same entity is not known a priori and often relies on machine learning techniques [11, 12, 13, 14].

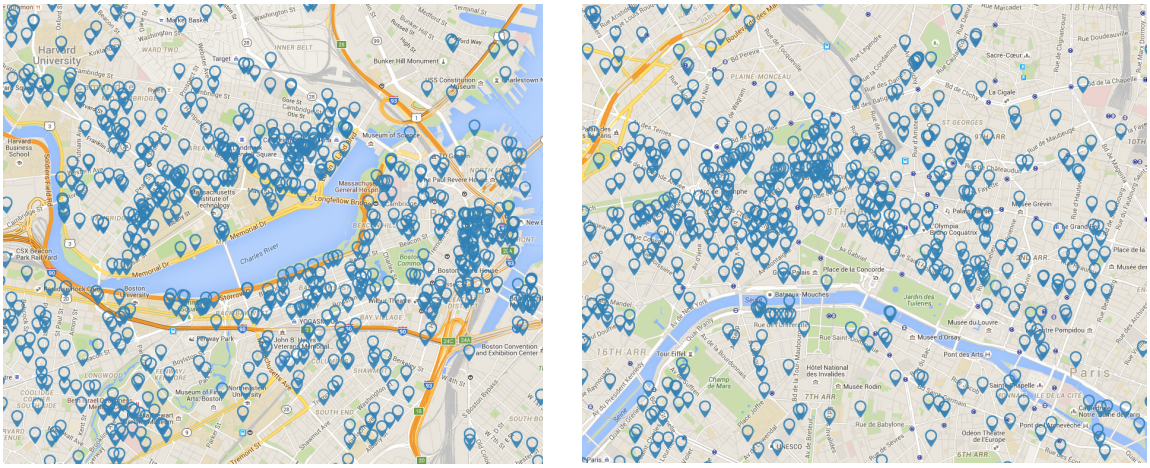

Figure 1: High-resolution geolocations of assignees in the Boston (left) and Paris (right) regions. All high-resolution geolocations in Boston are included ($\sim$ 1.1K within 10 km of the city center), but due to the density of assignees in Paris,only lat/longs that are seen on at least three patents are shown ($\sim$ 5.2K within 10km of the city center). Clusters of geolocations are observed near MIT, MGH, BU, and downtown Boston. Maps drawn using the tools of Ref. [15].

In this paper, we describe a straightforward but accurate approach to the disambiguation

problem using high precision geolocation of assignee and inventor addresses. Two inventors (or two assignees) that provide exactly the same high-resolution address and *also* have 'similar' names are very likely to refer to the same entity. Thus, knowing that two entities have exactly the same address allows a great deal of flexibility in name matching, and we design two simple string matching approaches to link similar names that share a high-resolution geolocation. Inventors and assignees that have addresses with low resolution, which are extremely common in the USPTO data, are linked to exact name matches nearby, greatly increasing the coverage of the disambiguation. Additional linking can be accomplished by searching for similar names that share other characteristics in common (such as co-inventors or co-citations). We show that this approach provides a complete disambiguation of inventors and assignees on 8.47 million patents, and that the precision and recall of the resulting disambiguation is superior to or competitive with other well known disambiguation methods.

Our results are a significant improvement over existing disambiguation techniques in a number of ways. This is the first simultaneous disambiguation of names in the EPO, PCT, and USPTO, providing a bridge between two existing but often non-overlapping streams of research focused on European and US patents. This combined disambiguation is of particular use in minimizing bias due to European or American institutions tending to apply for patents in their own domestic offices, and ensuring fair coverage of research hubs on both sides of the Atlantic. A disambiguation of both the assignee and inventor names is an aspect lacking in most freely available disambiguations of patent databases. This is of particular importance for understanding the flow of inventors between institutions (whether in the same region or in different countries), as well as identifying the institutional research hubs that are major players in each country. Finally, we emphasize that our disambiguation also incorporates high-resolution geolocations of all inventor and assignee, providing geospatial information that can go beyond traditional regional or national aggregations based on (often arbitrary) administrative boundaries. We believe the breadth and level of detail in the database produced by this work will be of great value to researchers, and will make the data freely available for noncommercial use.

2 Patent Data Sources

In this paper, we combine a number of distinct databases covering different patent offices: the US Patent and Trademark Office (USPTO), the European Patent Office (EPO), and patents filed under the Patent Cooperation Treaty (PCT). The USPTO and EPO generally provide protection of intellectual property within the US and European Union (EU) respectively, while the PCT provides a method of international protection for a patent filed domestically. In general, US and Japanese entities tend to be most represented in the USPTO, European entities in the EPO, and other nations (particularly developing nations)

in the PCT database.

Information about USPTO patents is extracted from Harvard's Dataverse project from the work of Ref.[12] (covering patents granted between 1975 and 2010), while information about EPO and PCT patents is extracted combining the Organization for Economic Cooperation and Development's (OECD) RegPat[16] and Citation[17] databases (covering patents applied for between 1977 and 2011). We also incorporate the OECD's Triadic Patent Families[18] database in this work, which identifies patents that cover the same technology filed in the USPTO, EPO, and Japanese Patent Offices (JPO). This combination of the USPTO, EPO, and PCT databases provides the application year for each patent, the names and addresses[1] of all inventors and assignees, citations between patents (including those filed in different offices), and patents that are found in the same triadic family.

This paper will focus heavily on direct collaborations of assignees and inventors, direct citations, or exact family correspondence, and will not use additional information available in the databases such as the technological classes of each patent. Technological classifications have multiple potential levels of aggregation, ranging from very general topics like "Pharmaceuticals" to specific topics like the full International Patent Classification (IPC) code "A61K003/121: Medicinal preparations containing acyclic ketones," and it is not clear which classification scheme is best for the disambiguation process.

Our method will be compared with two other freely available methods of name cleaning and disambiguation that exist in the literature: for inventor disambiguation we will compare our results to the output of Ref. [12], and for assignee disambiguation we will compare our results to the name cleaning provided by the OECD Harmonized Applicant Name [19] (HAN) database.

3 Geolocation and Disambiguation of Assignees and Inventors

The fundamental difficulty that must be overcome in name disambiguation is the possibility of alternate or error-ridden spellings of names or addresses in the database. Two examples of names requiring disambiguation are depicted in Fig. 2: on the left are some of the names associated with the National Institutes of Health (NIH) in Bethesda MD (assignee on $\sim 4,900$ patents), and some of the names associated with a prolific inventor, Rosa Maria Cuberes-Altisent, in the Barcelona area (inventor on 90 patents). In both cases, the subsets of names shown in the white boxes of Fig. 2 indicate the extreme heterogeneity in some

[1] Note there may be no legal requirements or error checking on these addresses. Inventors are free to use their home or office addresses, and assignees can use a PO Box if desired.

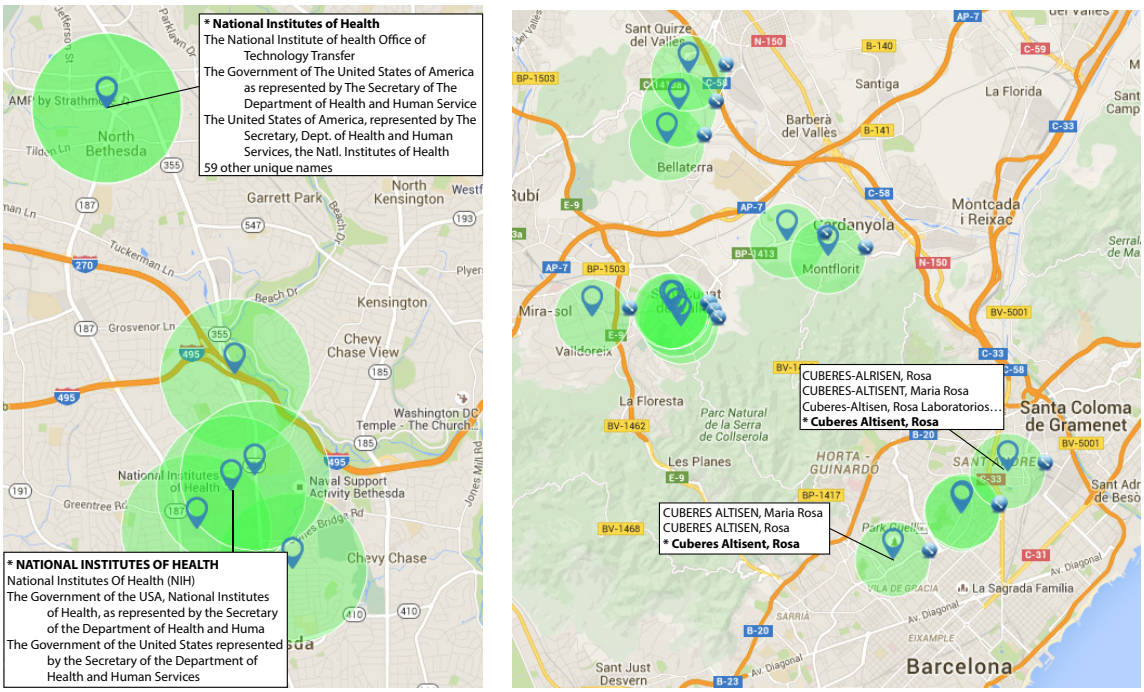

Figure 2: Geolocations of assignee addresses (left) or inventor addresses (right) for two examples: the NIH in Bethesda Maryland, and an inventor in the Barcelona area with many names and addresses. In both cases, there is a great deal of heterogeneity of names at some geolocations, but many of the names are 'similar' at precise addresses. There are also exact name matches nearby (within 20km in this paper), highlighted in bold text. Maps drawn using the tools of Ref. [15].

inventor or assignee names, with $\sim$ 90 unique alternate spellings for the NIH and $\sim$ 10 unique alternate spellings for Rosa Maria Cuberes-Altisent (removing spacing, punctuation, and capitalization). Disambiguation of these names requires not only matching all of the possible variations in the spelling of the institution or individual, but also *not* matching other names that refer to different entities with similar names. A wide range of methods of varying complexity have been generated to solve this disambiguation problem for authors of publications [11, 20, 21, 22, 23], patent assignees [24, 13, 25], and patent inventor names [12, 14, 26] (with the disambiguation of Li et. al [12] a recent and comprehensive result for the USPTO). In the case of inventor disambiguation, these methods will generally compare pairs of names using the similarity of the text of the names as well as data regarding the assignees, patent citations, patent classes, and geographical information.

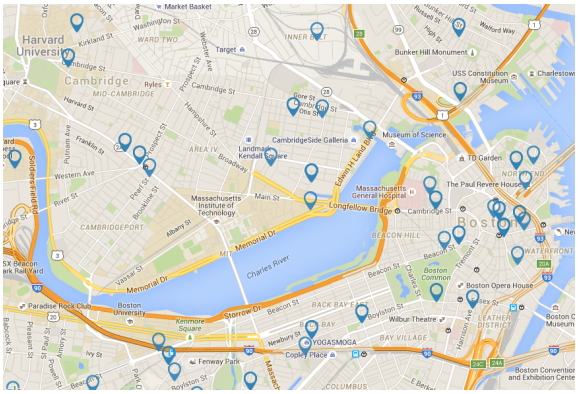

Figure 3: The handful of geolocations of low-resolution addresses in the Boston area. Note that there are $\sim$40K patents found at these $\sim$180 points, in comparison to the $\sim$15K patents found at the $\sim$1,100 points in Fig.1. The loss of specificity in these geolocations would make flexible name matching unreliable at these low resolution geolocations.

The geographical information found in the USPTO typically suffers from low quality addresses, where below 5% of USPTO inventors complete the street field in their address on the patent (city- or zipcode-level information is the highest resolution available). At this resolution, geolocation can be used as one of many rough indicators of the similarity between two names when comparing them for disambiguation. However, patents in the EPO or PCT databases are found to contain higher resolution addresses in a far greater fraction of cases (where a street number, street, city, state, and zip code are often all provided), which can provide much greater specificity when comparing inventors: if two inventors have 'similar' names *and* live at exactly the same address, it is far more likely they refer to the same person than if they had 'similar' names and lived in the same general area. The same state of affairs exists for assignee names, with the EPO and PCT addresses often having street-level information and the USPTO addresses tending to be of low quality.

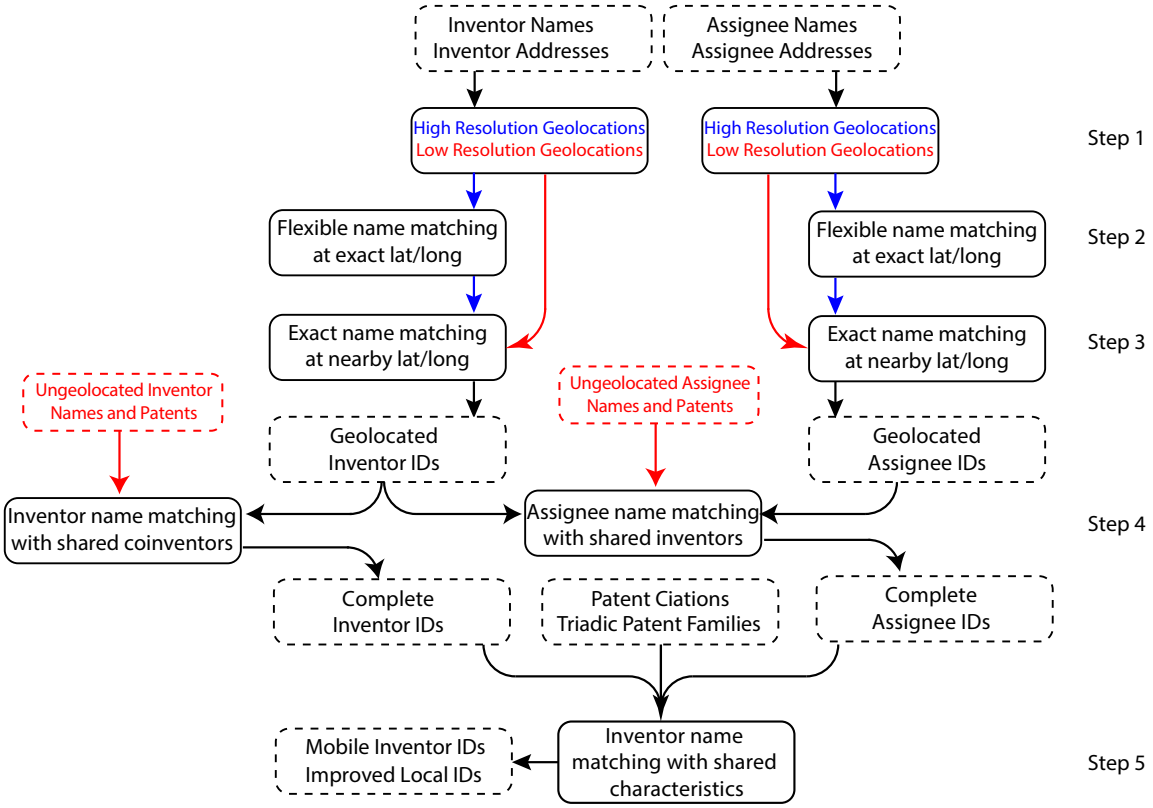

Figure 4: Schematic diagram of the method. Names and addresses are first geolocated (discussed in Sec. A.1), after which there is a search for 'similar' names at exact high resolution lat/longs (the meaning of 'similar' is discussed in Sec. A.2 for assignees and Sec. A.3 for inventors). Once similar names are clustered at each lat/long, 'nearby' exact name matches are linked (described in Sec. A.4). Assignees and inventors names that could not be geolocated are linked to disambiguated names with shared coinventorships (described in Sec. A.5). Finally, similar inventor names for inventors that are not found at the same address but share other characteristics in common before are linked, creating identifiers for mobile inventors (described in Sec. A.6).

Our strategy for disambiguation will be to leverage these high precision addresses in the EPO and PCT by flexibly matching names that are simultaneously found at exact, high-precision geolocations in any patent office, then to match 'nearby' names that are exactly the same. The general idea is sketched in Fig. 2 for specific assignee and inventor names, and a schematic of the methodology is diagrammed in Fig. 4. A detailed summary of each step can be found in the appendix. We first geolocate the assignees and inventor addresses for every patent in the three databases ($\sim$ 4 million unique addresses) using Yahoo's YQL API [27, 28], converting the text into likely latitude/longitude (lat/long) pairs and the quality of that geolocation (step 1 in Fig. 4). The quality returned by YQL[2] generally indicates [29] if the geolocation was resolved at the level of a point (e.g. "55 Fruit Street, Boston, MA, USA"), line (e.g. "Fruit Street, Boston, MA, USA"), zip code (e.g. "02114 MA USA"), city (e.g. "Boston, MA, USA"), state (e.g. "MA, USA"), or country (e.g. "USA").

For every high-quality geolocation in the data (those better than line-resolution) we attempt to flexibly match similar strings (step 2 in Fig. 4), with specific examples in Fig. 2: assignees with names involving 'institute' or 'health' at identical lat/longs are likely to be referring to the NIH, and inventors with names like 'cuberes' and 'rosa' at identical lat/longs are likely referring to Rosa Maria Cuberes-Altisent. The specifics of the matching can be found in Sec. A.3, but generally speaking we match pairs of assignee names that share either one rare or two common words (in any order, up to one spelling error) and match pairs of inventor names for which the first two words in either name occurs in the other (in any order, up to one spelling error), so long as both names share a high-resolution geolocation. After this first round of disambiguation, which only compares pairs of names found at exactly the same high-resolution geolocations, we search for exact name matches that were geolocated to a lat/long pair of low quality (having a YQL quality indicator on the level of city-resolution) in step 3 of Fig. 4. For example, we link would the names at the indicated high-resolution geolocations due to the simultaneous name matchings of "National Institutes of Health" and "Cuberes Altisent, Rosa" occurring in Fig.2. We also search for shared prefixes between names (where an inventor may leave his or her middle name out on some patents) to improve the linking of low resolution names. The process is fully described in Sec. A.4. Steps 2-3 in Fig. 4 are run in parallel to one another, with assignee and inventor disambiguation independent of each other up to this point.

Patents that have assignees that could not be geolocated at the precision of city-level or better are common, particularly in the USPTO data, and it is important to include patents

[2]In early 2016, Yahoo discontinued the use of the particular API used in this paper, and no longer appears to produce the same quality indicator (see https://developer.yahoo.com/boss/search/#pricing). Other geolocation APIs often provide quality indicators (e.g. Google labels geolocations as 'rooftop' for the highest resolution, with other indicators for lower resolution). While the results of this paper are strongly dependent on the division between high and low resolution geolocations, we expect our results to be robust to changes of the specific quality indicator used.

with extremely poor geolocations in the disambiguation as well. Matching assignee names to the disambiguation not straightforward if multiple disambiguated IDs have the same name as the un-geolocated assignee: for example, one must determine whether the assignee "General Electric Corporation" without an address provided (occurring over 18,500 times in the USPTO) is referring the same institution as the "General Electric Corporation" found in New York, Ohio, Massachusetts, or many additional potential geolocations found for that name. We identify the most likely disambiguated ID for linking an unlocated patent as that with a similar name and shared inventor or an inventor in close physical proximity to the geolocated assignee (a detailed description of this method is found in Sec. A.5).

The results of the disambiguation after step 4 of Fig. 4 produces assignee and inventor IDs that are highly localized, linking names only if their geolocations are within a 20km radius of one another. For assignees this localized partitioning is reasonable, with different corporate offices, research labs, or subsidiaries of the same company in different cities treated as independent entities (for example, "General Electric Corporation" in Schenectady NY is not the same as "General Electric Corporation" in Cincinnati OH). Inventors are fundamentally different in that they can be mobile, moving between different cities and countries. We therefore perform a final round of linking on the disambiguated inventor IDs in step 5 of Fig. 4 based on characteristics in common that indicate a potential relationship: a shared co-inventor, shared assignee, shared triadic family, or citation. Note that we do not include the technological classes of the patents in our measure of similarity, because they are less personal (working on a research topic is far less informative than collaborating with an individual) and the 'correct' level of aggregation is unclear. A link between disambiguated IDs is performed for inventors with identical names and one characteristic or similar names and two characteristics in common, providing a method to identify the movement of unique individuals between different cities.

This procedure produces a disambiguation of ∼9.3M patents in the three patent offices. A total of ∼804K unique geolocated assignee identifiers are found, as well as ∼443K unlocated assignee IDs (due to missing addresses that could not be linked to the disambiguation). We find ∼ 3.8M unique geolocated inventor IDs that do not move (i.e. are only found in the same 20km radius), ∼425K mobile inventors (IDs that are linked to geolocations more than 20km apart), and ∼ 439K inventor IDs that could not be linked to a location.

The algorithm uses geospatial information throughout, so the unlocated IDs are likely to be error prone and unreliable. Because correction of errors in spelling occurs only in steps 1 and 5 of Fig. 4, it may be important to further distinguish between 'high-quality' disambiguated IDs (those involving both geolocation and spelling correction), and 'low-quality' disambiguated IDs (those using only geolocations to link nearby identical names). High-quality inventor or assignee IDs can be linked to a high-resolution geolocation, while high-quality inventor IDs also include those linked in step 5 through additional shared

Assignee Disambiguation Statistics

	EPO	PCT	USPTO
Patents with Assignees	2.68	2.36	3.61
Patents with Assignee addresses	2.68 (100%)	2.36 (100%)	2.01 (56%)
Patents linked to geolocated IDs	2.62 (99%)	2.28 (97%)	3.18 (88%)
Patents with all high-qual disambiguated IDs	1.95 (73%)	1.44 (61%)	1.89 (52%)

Inventor Disambiguation Statistics

	EPO	PCT	USPTO
Patents with Inventors	2.68	2.34	4.24
Patents with Inventor addresses	2.67	2.27	4.24
Patents linked to geolocated IDs	2.62 (98%)	2.24 (96%)	4.15 (98%)
Patents with all high-qual disambiguated IDs	2.05 (78%)	1.34 (57%)	2.09 (49%)

Table 1: Summary of the assignee and inventor coverage of the disambiguation in the three patent offices (patent count in millions). Listed are the number of patents having an assignee or inventor listed in the raw data (note that these fields may be blank); the number of patents with any address; the number of patents for which all of those entities is linked to a geolocated and disambiguated ID (of either high or low quality); and the number of patents for which all of those entities are high-quality (the disambiguation involved both geospatial information as well as noise correction in the name).

characteristics. We find a total of ∼290k high quality assignees and ∼1.8M high quality inventor IDs using this definition, with the best coverage in the EPO and worst coverage in the USPTO (see Table 1). Our method produces a complete disambiguation of all assignees and inventors on ∼ 8.5M patents, and complete high-quality disambiguations on ∼ 4.3M patents worldwide (see Table 2). The disambiguation thus completely geolocates and disambiguates all names on over 90% of the patents in these offices, with nearly 50% of the those patent disambiguations incorporating high resolution addresses and error correction in the names.

	EPO	PCT	USPTO	tot
Patents in Database	2.68	2.36	4.04	9.08
Patents with all names located & disambiguated	2.58 (98%)	2.22 (98%)	3.67 (91%)	8.47 (93%)
Patents with all high-quality disambiguations	1.74 (66%)	1.08 (48%)	1.47 (36%)	4.29 (47%)

Table 2: Summary of the total coverage of the disambiguation in the three patent offices (patent count in millions). Listed are the number of patents having an assignee or inventor listed in the raw data (note that these fields may be blank); the number of patents with any address; the number of patents for which all of those entities is linked to a geolocated and disambiguated ID (of either high or low quality); and the number of patents for which all of those entities are high-quality (the disambiguation involved both geospatial information as well as noise correction in the name).

4 Evaluation of the Method

4.1 Splitting, Lumping, Precision, and Recall

We will quantify the compatibility of a trial disambiguation to the 'correct' ground truth partition using four different indicators: splitting, lumping, precision, and recall. Splitting and lumping are two related statistics measure the similarity between two partitions that have been previously used for inventors in the UPSTO[12], which respectively estimate the number of patents that are missing in the trial disambiguation in comparison to the benchmark (i.e. a true inventor is split into multiple IDs in the trial) and the number of patents that are added in the trial disambiguation compared to the benchmark (i.e. multiple trial IDs point to the same true inventor). The splitting associated with trial ID i is determined by identifying the true ID that shares the most patents, m_i; with $\mathbf{P}_i^t$ the set of patents invented by i in the trial partition and $\mathbf{P}_{m_i}^T$ the set of patents invented by m_i in the true disambiguation, the splitting is $s_i = |\mathbf{P}_i^t \backslash \mathbf{P}_{m_i}^T|$ (i.e., the number of patents in the trial that *are* associated with i but *not* with m_i in the true disambiguation). Lumping is similarly defined for the true ID j by identifying the trial ID M_j with which it shares the most patents, with the lumping of j given by $l_j = |\mathbf{P}_j^T \backslash \mathbf{P}_{M_j}^t|$ (the number of patents in the trial that *are* associated with M_j but *not* with j). To perform this matching, we identify the best link for the ID with the largest number of patents (i or j, depending on if we are computing splitting or lumping), then progressively search for the best matches for smaller IDs that have not been previously assigned. Rarely occurring IDs may not be assigned a best match using this procedure (if all potential matches were previously assigned), and in this case all occurrences of that ID will be treated as erroneous. The total splitting and lumping of the trial partition are given by

$$\text{split} = \frac{\sum_i s_i}{\sum_i |\mathbf{P}_i^t|} \qquad \text{lump} = \frac{\sum_j l_j}{\sum_j |\mathbf{P}_j^T|}, \tag{1}$$

which estimates the fraction of patent that suffer from a splitting or lumping error.

Splitting and lumping in Eq. 1 link trial IDs to true IDs that share the most patents in common, and treat patents assigned to a different ID as an error. While this is a useful statistic, it ignores the degree of splitting or lumping that has occurred: splitting remains unaltered by the number of IDs a true identity is associated with in the trial (and similarly for the lumping statistic), suggesting that some features of the similarity between partitions may be overlooked. We therefore also examine a pairwise measure of precision and recall of the trial and true disambiguations, schematically diagrammed in Fig. 5. The disambiguated name-to-ID's $\mathbf{n}^T(p) = \{n_i^T(p)\}$ for the ith name on patent p that can be compared to the disambiguated names in the trial disambiguation. For each pair of patents p_1 and p_2, we can determine the number of true positives, false positives, and false

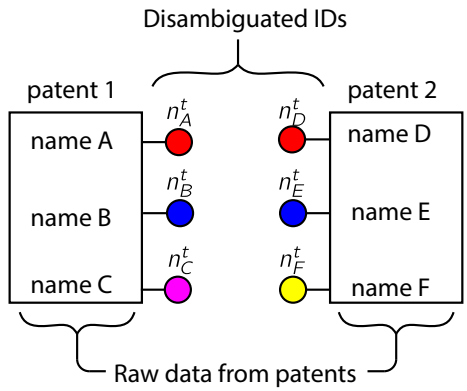

Figure 5: Schematic of the pairwise measurements of precision and recall. Each name is assigned a unique ID in the benchmark (indicated by the color of the circles) and a ID in the trial disambiguation (indicated by $\mathbf{n}^t$). Two true positives occur if $n_A^t = n_D^t$ and $n_B^t = n_E^t$. Two false negatives would occur if $n_A^t \neq n_D^t$ and $n_B^t \neq n_E^t$. A false positive occurs if $n_C^t = n_F^t$.

negatives between the trial and benchmark IDs by comparing the number of overlapping identifiers, as diagrammed in Fig. 5. Defining $x_T(p_1, p_2) = |\{n_k^T(p_1) \cap n_k^T(p_2)|$ as the size of the intersection of the IDs in the benchmark and $x_t(p_1, p_2) = |\{n_k^t(p_1) \cap n_k^t(p_2)|$ the size of the intersection in the trial disambiguation, there are at most $TP(p_1, p_2) = \min(x_T, x_t)$ IDs that agree in both partitions, $FN(p_1, p_2) = \max(0, x_T - x_t)$ matches in the benchmark not seen in the trial, and $FP(p_1, p_2) = \max(0, x_t - x_T)$ matches in the trial partition that don't match in the benchmark. An estimate[3] of the precision and recall for the trial is then

$$\text{prec} = \frac{\sum_{p_1 \neq p_2} TP(p_1, p_2)}{\sum_{p_1 \neq p_2} [TP(p_1, p_2) + FP(p_1, p_2)]} \qquad \text{rec} = \frac{\sum_{p_1 \neq p_2} TP(p_1, p_2)}{\sum_{p_1 \neq p_2} [TP(p_1, p_2) + FN(p_1, p_2)]} \quad (2)$$

Using the definitions in Eq. 2, if all matches found in the trial are also found in the benchmark, the trial disambiguation will have high precision; and if all matches in the benchmark are found in the trial, the trial disambiguation will have high recall. These definitions of precision and recall are independent of a matching between the two partitions (that is, associating each trial ID with a 'best' true ID), comparing all assignments on an equal footing (but at greater computational cost due to the pairwise comparisons required).

[3]Note that this method does not compare the positions of the name matches (only the number of matches), and thus neglects the possibility of a transposition of the IDs in the trial (e.g. if names A and E are incorrectly linked together and *simultaneously* names B and D are incorrectly linked together). Due to the large number of IDs for both assignees and inventors, this type of error is expected to have a negligible effect. This approximation can be relaxed, at an increased computational cost.

Gold Standard Inventor Disambiguation

All patents

	none	lower	upper	ours
# pats	1582	1320	1320	1321
# IDs	176	118	121	150
lumping (lower better)	29.8%	4.9%	5.3%	10.5%
splitting (lower better)	15.2%	3.8%	4.2%	9.5%
precision (higher better)	0.50	0.97	0.97	0.94
recall (higher better)	0.36	0.88	0.88	0.82

Only high-quality IDs

	none	lower	upper	ours
# pats	825	606	606	608
# IDs	88	58	59	62
lumping (lower better)	30.4%	2.0%	2.6%	4.1%
splitting (lower better)	10.5%	1.5%	2.0%	4.1%
precision (higher better)	0.22	0.97	0.97	0.98
recall (higher better)	0.23	0.97	0.96	0.92

Table 3: Benchmarking of inventor disambiguations on the Golden Standard benchmark. The column 'none' refers to a name disambiguation that simply removes punctuation and differences between upper and lowercase letters, the columns 'lower' and 'upper' refer to the two disambiguations of Li et al[12], and the column 'ours' refers to this work. In the table on the left, the benchmark includes all patents found in the Golden Standard as well as all USPTO patents in the disambiguations with at least one trial ID (thus including patents erroneously assigned to the disambiguated IDs). In the table on the right, the dataset is restricted only to patents for which our disambiguation has all 'high-quality' IDs, for which both geolocation and spelling errors were potentially corrected.

Harvard Inventor Disambiguation

All patents

	none	lower	upper	ours
# pats	587	587	587	587
# IDs	877	827	829	827
lumping (lower better)	4.0%	0.7%	0.7%	1.6%
splitting (lower better)	3.6%	0.7%	0.9%	1.5%
precision (higher better)	1.00	1.00	1.00	1.00
recall (higher better)	0.89	0.96	0.96	0.96

Only high-quality IDs

	none	lower	upper	ours
# pats	115	115	115	115
# IDs	177	165	165	164
lumping (lower better)	5.6%	0.2%	0.3%	0.0%
splitting (lower better)	4.5%	0.3%	0.3%	0.0%
precision (higher better)	1.00	1.00	1.00	1.00
recall (higher better)	0.86	0.99	0.99	1.00

Table 4: Benchmarking of inventor disambiguations on patents assigned to Harvard. The columns are the same as in Table 3.

In order to evaluate the accuracy of our inventor disambiguation, it is necessary to find a set of ground truth identification of patent inventors: a manually curated subset of the patent data for which a correct disambiguation of the names has been performed. Such a benchmark exists for ~ 100 USPTO inventors of ~ 1300 patents in the area of engineering and biochemistry that has been used as a golden standard in the literature[12]. This is a complete disambiguation of all patents invented by a specific set of inventors. We also generated a manual disambiguation of USPTO patents filed by assignees containing the phrases "harvard college" and "harvard university" (case insensitive) in the name. This second benchmark is a complete disambiguation of all inventors on a specific set of patents, with 587 patents with 1,000 name/address pairs and 827 disambiguated inventor IDs (using our approach). In Tables 3-4, we compare the splitting, lumping, precision, and recall of four potential disambiguations: 'none' (meaning that case and punctuation are ignored but no further disambiguation is performed), 'lower' and 'upper' (two outputs of the disambiguation of Ref[12]), and our approach. For the golden standard benchmark (Table 3), all patents linked to a trial ID are included when computing the statistics (including incorrectly-assigned patents, representing false positives), but un-disambiguated collaborators of these inventors are ignored. For the Harvard inventor disambiguation shown in Table 4, only the disambiguated patents are included (and no additional patents that may represent false positives are added) because we do not have a complete list of the 'correct' patents for each inventor. The exclusion of such false positives increase the apparent precision of each method in Table 4, since exact name matches for inventors affiliated with Harvard are very likely referring to the same person.

In the left side of Table 3 and 4, we see that all disambiguation methods ('lower,', 'upper,' and 'ours') provide an improvement over the un-disambiguated results in all statistics, and in general the results of Li et. al. [12] produce a marginally higher improvement in all statistics in comparison to our approach. The quality of our disambiguation is comparable to that of Li's in the case of the Harvard dataset, but in the case of the Golden Standard benchmark our results have about half the lumping of the un-disambiguated name and major improvements in both precision and recall. The disambiguation of Li is a more significant improvement over the raw names in lumping, splitting, and recall, but is comparable to our approach in precision. A higher recall can be found by removing all IDs associated with low-quality disambiguations in our approach (those IDs that did not incorporate correction of spelling errors in steps 1 and 5 of Fig. 4). On the right sides of Tables 3 and 4, we discard all patents having at least one low-quality ID, as well as all names or Li IDs occurring on those patents. We find a significant improvement in all quality statistics using our approach in both benchmarks, although still marginally underperforming Li et al in both datasets. These results indicate that our inventor disambiguation is expected to have high precision in all cases (i.e. will have very few false positives), but higher recall (very few false negatives) for high-quality disambiguated names. Low-quality disambiguations are expected to underestimate the number of patents invented by the person, but not

significantly overestimate them.

Boston Area Assignees

	none	HAN	ours
# pats	22527	22175	22527
# IDs	376	691	357
lumping (lower better)	19.0%	11.5%	4.0%
splitting (lower better)	15.5%	12.8%	3.5%
precision (higher better)	1.00	0.99	1.00
recall (higher better)	0.74	0.76	0.99

Paris Area Assignees

	none	HAN	ours
# pats	18876	18011	18876
# IDs	279	480	248
lumping (lower better)	34.0%	23.8%	8.0%
splitting (lower better)	26.4%	25.3%	8.0%
precision (higher better)	1.00	0.97	0.97
recall (higher better)	0.51	0.53	0.95

Table 5: Benchmarking of assignee disambiguations in the EPO and under the PCT for assignees geolocated in the Boston (left) or Paris (right) areas. The column 'none' refers to a name disambiguation that simply removes punctuation and differences between upper and lowercase letters, the column 'HAN' refers to OECD's Harmonized Assignee Name identifier, and the column 'ours' refers to this work.

We are not aware of a freely available ground truth disambiguation of assignees, and in order to test the accuracy of our methods we manually generated our own benchmark from a by-hand disambiguation of a subset of patent assignees in the EPO and PCT data. To create a benchmark of manageable size, we focused on assignees in the EPO and PCT in specific regions active in specific fields of research. The OECD REGPAT database [16] provides geolocation information on the level of NUTS3, and we generated a list of all assignees on patents assigned to names in Boston or Paris[4]. These names include both regional assignees as well as external collaborators of those regional institutions, and in order to reduce the number of names to disambiguate, we retained only assignees names found on at least ten biopharma patents[5]. Exclusion of less-common names greatly reduced the number of manual matches that were required, but may introduce a bias against small startups, individuals that were assignees, or large institutions that rarely work in the biopharma fields. This produced a final set of ∼700 disambiguated assignees (from ∼900 raw names) on ∼23K biopharma patents in the Boston area and 640 disambiguated assignees (from ∼1,100 raw names) on ∼19K biopharma patents in the Paris area.

[4]The selection of assignees was made where Boston assignees have addresses in the NUTS3's US25017, US25025, or US25021 and Paris have addresses in NUTS3's beginning with "FR10" in the OECD RegPat database[16]. We note that the OECD's geolocation process is independent from our own, so this benchmark is not affected by any potential errors or incompleteness in our geolocation procedure.

[5]Biopharma patents are defined as having an IPC classification which falls under the fields "Pharmaceuticals" or "Biotechnology" using the WIPO field-level aggregation of patent classes [30]

In Table 5, we determine the error statistics for undisambiguated names; the HAN name harmonization[19], which accounts for synonyms such as "Co" and "Company" in assignee names; and the methods in this paper. We are not aware of another freely available assignee disambiguation of EPO and PCT patents that can be used for comparison. As was the case in Table 4, false positives from patents not found in the manual disambiguation are excluded from this trial, and the precision is likely to be inflated[6]. In all cases the HAN harmonization is a marginal improvement over the raw names, but our method produces a significant improvement over both in every statistic. The lower overall recall in the Paris area is due to many alternate spellings from the presence or absence of accents on words, a complication not typically present in the Boston area. Note that the names in the Paris area involve terms in French (e.g. "societé aononyme" is a commonly occurring substring of assignee names), and the high precision of our disambiguation in the Paris area suggests that the algorithm is robust to variations in language. Over 95% of the patents in Table 5 are assigned to solely high-quality assignee disambiguations, and there is little benefit in dividing the results into differing qualities of disambiguation (as was beneficial in Tables 3 and 4). The error statistics in Table 5 suggest that our assignee disambiguation is likely reliable in many major global research hubs.

5 Leading Inventors and Assignees

Bibliometric data has been implemented in a variety of contexts in both publications and patents[9, 8, 31, 1] to track individual careers, institutional collaborations, and regional growth. In almost all studies, major players represented by top authors, inventors, or institutions tend to be emphasized (as measured by the number of patents or publications, number of citations, or other quality indicators), due to their high productivity and output. It is therefore worthwhile to determine how accurately our approach is able to identify the top players worldwide, and what advantages our method may have over existing datasets.

In Table 6, we list the top ten players as determined by Li's 'lower' inventor disambiguation of USPTO inventors (the 'upper' disambiguation does not differ significantly on these names) compared to the output of our methods. We find a negligible difference (typically less than 2% variation in the patent counts) between the approaches in the majority of cases, but significant differences do sometimes occur. In the case of Gurtej Sandhu, we find our algorithm detects an additional ∼100 patents, an improvement over Li's approach sufficient to change his ranking from 7th to 5th place in our ordering. Our method sometimes does underestimate the patents invented by top players, particularly for Japanese inventors.

[6] Because our approach permits a maximum distance between geolocations of 20km on the disambiguation of assignees, it is unlikely that the precision using our approach is significantly higher than what would be observed if all patents were included.

This is due to the low-quality of geolocations provided by YQL in Japan: of the $\sim$ 100K address containing the word 'Tokyo,' 0.6% produced a high-resolution geolocation (meaning that virtually no Japanese inventor will have a high-resolution geolocation) but 82% are found on the neighborhood level or better (meaning they contained more information than simply a city name). It may therefore be reasonable to expect a higher number of splitting errors for Japanese inventors using our data. Table 6 indicates that, with the exception of these splitting errors of Japanese inventors, our method is able to reliably detect prolific inventors consistent with Ref. [12].

A surprising aspect of Table 6 is that not a single European is found in the list of the top 20 most prolific inventors. This is due to an intrinsic bias in the choice of USPTO patents: Europeans may be more likely to apply for patents in the EPO (which grants protection in Europe where they live and work) than in the USPTO. Non-US and non-European inventors and assignees may likewise be more likely to submit patents to under the PCT than in either the USPTO or EPO, to ensure international and domestic protection of their inventions. A significant advantage of our method is that these three major offices are linked in the disambiguation, permitting a more even accounting of highly prolific inventors. In Table 7, we see that while the top five inventors worldwide all focus their patenting activity most heavily in the USPTO, many prolific inventors focus their patent applications in the EPO or under the PCT. We are able to identify a number of top inventors from China (filing their patents almost exclusively under the PCT) and Germany (filing their patents in both the EPO and USPTO, typically), many of whom would not be seen in the top hundred inventors when focused solely on the USPTO. Table 7 suggests that our inventor disambiguation may more appropriate in studying the effects of global knowledge production and international knowledge spillovers through collaborations across the Atlantic as well as in emerging economies.

Another significant advantage of our methodology is the simultaneous disambiguation of assignees, allowing us to determine the major research institutions in addition to prolific inventors. In Table 8, we see the top patent assignees are dominated by computer and communications companies, with all members of the list firms with highly recognizable names. We also list the number of unique names that were involved in the disambiguation (spacing and punctuation ignored), highlighting the extreme variability of names that may occur. US and Japanese firms tend to patent in the USPTO (as was the case for inventors), and most European countries tend to patent more heavily in the EPO. The exceptions of Nokia Siemens in Finland and Ericsson in Sweden, who tend to evenly split their applications between the EPO and under the PCT, may be explained by these countries joining the European Union in 1995[8]. Note also that Nokia Siemens is found twice in this list: once

[7] "Gurtej Singh Sandhu" and "Gurtej S Sandhu" have different IDs in Li's disambiguation, but the same ID in ours.

[8] Note that while Sweden joined the EU in 1995, it was a signatory of the European Patent Convention in 1978.

USPTO rank (Li)	Name	Country	USPTO (Li[12])	USPTO (ours)	rank (ours)	% diff (high qual)	% diff (low qual)
1	Kia Silverbrook	AU	3570	3560	1	-0.3%	
2*	Shunpei Yamazaki	JP	2519	2398	2		-4.8%
3	Donald Weder	US (IL)	997	998	3	+0.1%	
4	Leonard Forbes	US (OR)	932	932	4	0%	
5	Paul Lapstun	AU	866	871	5	+0.5%	
6	Chang Hwan Hwang	KR	840	838	7	-0.2%	
7	Gurtej Sandhu[7]	US (ID)	782	869	6	+11%	
8	Warren Farnworth	US (ID)	734	723	8	-1.5%	
9	Salman Akram	US (ID)	676	676	9	0%	
10*	Jun Koyama	JP	671	533	15		-21%
11	William Wood	US (CA)	647	649	10	+0.3%	
12	Austin Gurney	US (CA)	620	620	11	0%	
13	Audrey Goddard	US (CA)	606	606	12	0%	
14*	Akira Suzuki	JP	605	377	43		-38%
15	Paul Godowski	US (CA)	587	587	13	0%	
16	George Spector	US (NY)	556	553	14	-0.5%	
17	Mark Gardner	US (TX)	521	516	16	-1.0%	
18	Simon Walmsley	AU	511	512	17	+0.2%	
19	Jay Walker	US (CT)	509	510	18	+0.2%	
20*	Tetsujiro Kondo	JP	509	385	38		-24%

Table 6: The top ten patent inventors by number of patents in the USPTO (1980-2010), ranked using the disambiguation of Li et al[12] and ours (matching IDs between the two algorithms were based on the number of shared patents). In the majority of cases our algorithm produces a nearly identical number of patents, but it performs worse on Japanese inventors that have many low quality geolocations. Inventor locations indicate the position of the majority of their patents, as most inventors provide addresses in multiple countries. The '% diff' columns refer to the percent change in the number of patents assigned to that identity in our disambiguation relative to Li's, and are divided between high quality (having both geolocation and spelling correction) and low quality (having only geolocation). IDs that are categorized as low quality using our approach are denoted by a *.

Total rank	Total pats	name	country	USPTO rank	USPTO pats	EPO rank	EPO pats	PCT rank	PCT pats
1	4623	Kia Silverbrook	AU	1	**3560**	2	523	3	540
2*	2503	Shunpei Yamazaki	JP	2	**2398**	-	83	-	22
3	1261	Paul Lapstun	AU	5	**871**	102	189	62	201
4	1260	Tadahiro Ohmi	JP	20	**472**	11	356	6	432
5	1244	Eberhard Ammermann	DE	25	447	1	**541**	33	256
6	1085	Donald Weder	US (IL)	3	**998**	-	44	-	43
7*	1075	Yi Xie	CN	-	11	-	3	2	**1061**
8*	1075	Yumin Mao		-	12	-	2	1	**1061**
9	1067	Craig Rosen	US (MD)	81	298	9	378	10	**391**
10	1052	Steven Ruben	US (MD)	-	245	6	390	7	**417**
11	1024	Gurtej Sandhu	US (ID)	6	**869**	-	57	-	98
12	1019	Leonard Forbes	US (OR)	4	**932**	-	37	-	50
13	1014	William Wood	US (CA)	10	**649**	48	244	-	121
14	977	Austin Gurney	US (CA)	11	**620**	53	236	-	121
15	955	Heinz Focke	DE	24	**448**	3	441	-	66
16	906	Audrey Goddard	US (CA)	12	**606**	82	206	-	94
17	897	Siegfried Strathmann	DE	-	176	10	**372**	14	349
18*	870	Akira Suzuki	JP	43	**377**	13	331	-	162
19	850	Chang Hwan Hwang	KR	7	**838**	-	12	-	0
20	846	Paul Dent	US (NC)	60	**331**	30	269	38	246

Table 7: The twenty most prolific inventors in the combined USPTO, EPO, and PCT according to our disambiguation (recall that Table 6 included only USPTO patents), along with the specific counts per office. Inventors rankings in each office are suppressed if they are not in the top 100 most prolific inventors in that office. The most common office to which each inventor applies is in bold face text. IDs that are categorized as low quality using our approach are denoted by a *.

in Germany and once in Finland, with the appropriate designations of a limited liability corporation ("GMBH" and "OY", respectively) in their countries. Our method correctly treats these as distinct entities despite their similar names, due to the 20km geolocation restriction on assignees. Note that in order to determine the number of patents held by a multinational corporation (in the case of Nokia Siemens, the NOKIA group), an ownership tree must be determined from some alternate datasource. All of the European firms in Table8 would see their ranking drop significantly if only the USPTO were taken into account, with Nokia-Siemens' German office the only EU research lab that would be included in the top 20. As was the case for inventors, by including the USPTO, EPO, and PCT, a broader, more international picture of major research institutions can be extracted than if only one were included.

Total rank	Total pats	name	country	# names	USPTO rank	USPTO pats	EPO rank	EPO pats	PCT rank	PCT pats
1	73586	IBM	US (NY)	20	1	**53027**	7	15320	23	5239
2	66518	Nokia Siemens	DE	30	15	14439	1	**33857**	2	18222
3	65356	Phillips	NL	13	25	9959	2	**32084**	1	23313
4	56481	Cannon	JP	1	2	**38483**	6	15359	54	2639
5	49288	Robert Bosch	DE	12	21	11979	3	**20624**	3	16685
6	45634	Matsushita Electric	JP	5	4	**24867**	8	14401	17	6366
7	44663	Sony	JP	3	5	**23783**	5	15626	22	5254
8	39147	NEC	JP	18	7	**22739**	14	10793	21	5615
9	38759	Samsung	KR	18	3	**2592**2	12	10912	80	1925
10	37306	BASF	DE	10	28	9448	4	**17475**	7	10383
11	34842	Hitachi	JP	5	6	**23661**	24	7832	37	3349
12	34225	GE	US (NY)	13	10	**18633**	9	12306	38	3286
13	31340	Eastman Kodak	US (NY)	5	12	**16672**	16	10628	28	4040
14	30076	Nokia Siemens	FI	26	39	7554	13	10871	6	**11651**
15	28194	Proctor & Gamble	US (OH)	18	34	8401	17	**10510**	9	9283
16	27870	Ericsson	SE	16	64	5118	15	10662	5	**12090**
17	27837	Intel	US (CA)	9	9	**18688**	62	3506	20	5643
18	27442	Du Pont	US (DE)	20	23	**10757**	19	9601	13	7084
19	26309	Toshiba	JP	2	18	**12342**	10	11322	53	2645
20	26128	Microsoft	US (WA)	6	13	**16629**	35	5384	26	4115

Table 8: The twenty largest assignee headquarters as measured by the total number of patents worldwide, 1980-2010, along with the nation of the geolocation and the number of names for that assignee that were used in the disambiguation. Also shown are the patent counts in the various offices as well as their rank in those offices. Many major assignees focus their patenting applications on the EPO or under the PCT, and a focus solely on USPTO would overlook major players in the global innovation network.

6 Conclusions

In this paper, we have described a new algorithm for disambiguating the assignee and inventor names for a significant fraction of the patent data from the EPO, PCT, and USPTO simultaneously. Our approach focuses heavily on high-resolution geolocation of assignee and inventor addresses, determined by uploading all of the address information provided in the databases to the Yahoo Query Language [27]. This conversion from text to latitude/longitude pairs not only provides the possibility of precisely locating inventors and assignees on a map, but also acts as a disambiguation of the (often error prone) address fields. We use the high-resolution addresses to search for similar names at the same address using a very flexible string matching algorithm that are geolocated to *identical* latitude/longitude pairs, then search for *identical* names within a radius of 20km. This formed a seed disambiguation for which un-geolocated patents were linked and inventor mobility was detected, producing a complete disambiguation of all assignees and inventors for over 90% of the patents worldwide, and complete high-quality disambiguations (those involving geolocation and spelling correction) on nearly 50% of all patents. We show that our method is able to accurately disambiguate both assignees and inventors on two benchmarks using a manual disambiguation of assignee names in the Boston and Paris areas for patents in the EPO and PCT, a widely used benchmark of ~ 100 inventors [12] from the USPTO, and a manual disambiguation of USPTO patents with Harvard as an assignee. We found our disambiguation produced significant improvement over the use of raw assignee and inventor names, although the refined disambiguation of USPTO inventors in Ref [12] tends to produce somewhat fewer errors than our approach. Our approach has significant advantages over existing patent disambiguations due to its extremely broad coverage of the USPTO, EPO, and PCT and the simultaneous disambiguation of both assignees and inventors. This dataset provides a unique platform to study inter-institutional collaborations, inventor career trajectories, and the evolution of combinations of differing types of knowledge in new innovations. Our disambiguation has relied heavily on the geolocation of assignees and inventors, and this high-resolution geospatial information can be further leveraged for valuable studies as well, including inventor mobility between cites (without reference to arbitrary administrative boundaries) and spatial effects on knowledge spillovers. While we will continue refining our disambiguation and applying it to these topics, we intend to release this data to enable the scientific community to use these results as broadly as possible.

7 Data Release

Our data is being made freely available for noncommercial use, and will be released after a successful peer review evaluation. The data will be located at

https://github.com/gmorriso/PatentDisambigData
when available. The data that will be provided includes

- A patent-to-entity list of all 8.5 million patents to the unique assignee and inventor IDs, for which all entities can be linked to a high-resolution address.
- A correspondence between each assignee and inventor ID and the various names that were disambiguated using that ID.
- The manual institution disambiguation of assignees in the Boston and Paris areas, and the manual disambiguation of Harvard inventors for benchmarking.

For questions prior to release of the data, please contact greg.morrison@imtlucca.it.

Acknowledgements: We acknowledge support from the Crisis Lab Project (MIUR, Italy) and from the CERM Foundation (Switch Project). We are also grateful for the assistance of Anna Horodok on some of the manual disambiguation and for many useful conversations with Orion Penner.

References

[1] J Owen-Smith and WW Powell. Knowledge networks as channels and conduits: The effects of spillovers in the boston biotechnology community. *Org Sci*, 15:5, 2004.

[2] E Miguélez, , and R Moreno. Knowledge flows and the absorptive capacity of regions. *Research Policy*, 44:4, 2015.

[3] M Marx, D Strumsky, and L Fleming. *Noncompetes and inventor mobility: Specialists, stars, and the Michigan experiment.* Division of Research, Harvard Business School, 2007.

[4] Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. Foreign born scientists: Mobility patterns for sixteen countries. Technical report, National Bureau of Economic Research, 2012.

[5] A Chessa, A Morescalchi, F Pammolli, O Penner, AM Petersen, and M Riccaboni. Is europe evolving toward an integrated research area? *Science*, 339:650, 2013.

[6] C Franzoni, G Scellato, and P Stephan. Foreign born scientists: Mobility patterns for sixteen countries. *Nature Biotech*, 30:1250, 2012.

[7] M Riccaboni, WW Powell, F Pammolli, and J Owen-Smith. Public research and industrial innovation: A comparison of us and european innovation systems in the life sciences. *Science and Innovation. Rethinking the Rationales for Funding and Governance. Cheltenham: Edward Elgar*, pages 169–201, 2003.

[8] Andrea Morescalchi, Fabio Pammolli, Orion Penner, Alexander M Petersen, and Massimo Riccaboni. The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy*, 44(3):651–668, 2015.

[9] Greg Morrison, Eleftherios Giovanis, Fabio Pammolli, and Massimo Riccaboni. Border sensitive centrality in global patent citation networks. *Journal of Complex Networks*, 2(4):518–536, 2014.

[10] H Youn, D Strumsky, LMA Bettencourt, and J Lobo. Invention as a combinatorial process: evidence from us patents. *J of The Royal Soc Interface*, 12:20150272, 2015.

[11] VI Torvik and NR Smalheiser. Author name disambiguation in medline. *ACM Trans Knowl Discov Data*, 3:11, 2009.

[12] GC Li et al. Disambiguation and co-authorship networks of the us patent inventor database (19752010). *Research Policy*, 43:941, 2014.

[13] P Cuxac, JC Lamirel, and V Bonvallot. Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 97:47, 2013.

[14] Michele Pezzoni, Francesco Lissoni, and Gianluca Tarasconi. How to kill inventors: testing the massacrator© algorithm for inventor disambiguation. *Scientometrics*, 101(1):477–504, 2014.

[15] Free Map Tools. *http://www.freemaptools.com/radius-around-point.htm*, 2015.

[16] OECD REGPAT database. July 2014.

[17] OECD Citation database. July 2014.

[18] OECD triadic patents families database. July 2014.

[19] OECD HAN database. July 2014.

[20] H Han, L Giles, H Zha, C Li, and K Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004*, page 296, 2004.

[21] L Tang and J Walsh. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84:763, 2010.

[22] IS Kang et al. On co-authorship for author disambiguation. *Inf. Process. and Manag.*, 45:84, 2009.

[23] M Levin, S Krawczyk, S Bethard, and D Jurafsky. Citation-based bootstrapping for large-scale author disambiguation. *J Am Soc Info Sci and Tech*, 63:1030, 2012.

[24] B Balsmeier et al. Automated disambiguation of us patent grants and applications. *http://tinyurl.com/BBalsmeier*, 2015.

[25] BH Hall, AB Jaffe, and M Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. Technical report, National Bureau of Economic Research, 2001.

[26] SL Ventura, R Nugent, and ERH Fuchs. Methods matter: Revamping inventor disambiguation algorithms with classification models and labeled inventor records. *SSRN eLibrary*, 2012.

[27] Yahoo Inc. YQL web service URLs. *https://developer.yahoo.com/yql/guide/yql_url.html*, 2015.

[28] Roberto Catini, Dmytro Karamshuk, Orion Penner, and Massimo Riccaboni. Identifying geographic clusters: A network analytic approach. *Research Policy*, 44(9):1749–1762, 2015.

[29] Yahoo Inc. Supported response formats. *https://developer.yahoo.com/boss/geo/docs/supported_responses.*
quality, 2015.

[30] U Schmoch. Concept of a technology classification for country comparisons. *Final Report to the World Intellectial Property Office (WIPO)*, 2008.

[31] Jason Owen-Smith, Massimo Riccaboni, Fabio Pammolli, and Walter W Powell. A comparison of us and european university-industry relations in the life sciences. *Management science*, 48(1):24–43, 2002.

[32] OECD REGPAT database. January 2014.

[33] B Quinion. Open source search based on OpenStreetMap data. *https://github.com/twain47/Nominatim*, 2015.

[34] Google Inc. The google maps geocoding API. *https://developers.google.com/maps/documentation/geocoding/intro*, 2015.

[35] GeoNames. *http://www.geonames.org*, 2013.

A Methodology

A.1 Geolocation

Our disambiguation relies heavily on the geolocation of inventor and assignees using the addresses provided in the patent data. The names and addresses for each patent were extracted from the OECD Regpat January 2014 database [32] for the EPO and PCT patents, and from the patent database provided by Li et al [12] for the USPTO patents. The addresses are uploaded to yahoo's YQL [27] API (with a UTF-8 encoding of the

address; no additional cleaning performed), with a JSON response returned by the server containing a great deal of information about the geolocation(s) of that address. If the address was successfully located by YQL, we extracted all latitude and longitude data along with the quality of that geolocation [29], where the quality is an assessment of how precise the lat/long is (e.g. street level vs. city level). Note that in principle other geolocation APIs could be used (e.g. OpenStreetMap [33] or Google places [34]), but YQL was chosen due to its familiarity and ease of use. After the geolocation was complete on the January data, we acquired the July 2014 OECD Regpat database [16]. The geolcations extracted from the January database addresses were applied to the July database without modification, so some addresses from the newer data release may be missing.

The number of patents with geolocations are listed in Table A.1, with high coverage of the geolocation for all inventors and all assignees *except* for the USPTO. 1,847,909 USPTO patents have absolutely no assignee address information (no assignee provided, or no information about that assignee listed). Only 2,168,220 USPTO patents have assignee address information of any kind in the database (and ∼2M have address at the resolution of city or better). As our algorithm depends strongly on geolocation, only these ∼2M USPTO patents will have even a chance of acquiring a disambiguated assignee. Over 95% of the patents filed in each office have at least one inventor geolocation, and coverage is good in the EPO and PCT for assignees as well. Only ∼90% of the assignee addresses for the USPTO result in a geolocation, due to the relatively low quality of assignee addresses found in the data.

	EPO	PCT	USPTO
Patents in the database	2.67	2.37	4.14
Patents with inventor geolocation(s)	2.54	2.18	4.06
Patents with assignee geolocation(s)	2.58	2.24	1.95

Table 9: Patents in the database (numbers in millions of patents), and the address information included in them. Here, 'geolocations' refer to address information uploaded to YQL that returned at least one non-empty response. Empty responses are likely due to missing addresses (particularly due to the USPTO, where many assignees addresses are not provided).

A.2 Assignee name matching

Assignee disambiguation can be extremely difficult due to the large number of alternate spellings of assignee names. An example of a difficult disambiguation is shown in Table A.2 for alternate naming for the National Institutes of Health (NIH) in the Rockville / Bethesda areas of Maryland, USA, which is an agency of the Dept. of Health and Human services. Patents produced by the NIH may have assignee names that solely include references to

the Department to which it reports, or completely exclude the Department, or mention the Department in conjunction with the Institutes. While this extreme variability makes disambiguation difficult, we note that each of the addresses in this table are geolocated to the same lat/long with high quality using YQL. The geolocation is thus providing two services in this respect: in addition to locating the specific place where the institution is located, it is providing a robust disambiguation of the addresses. The similarity between the names in Table A.2 in conjunction with their precise geolocation to *identical* points certainly suggests that these names likely refer to the same entity, and in general one has good reason to believe that 'similar names' found at identical high precision street address are likely to refer to the same institution. Our algorithm for the first step of disambiguation is thus to block all names according to their high-resolution geolocations and search for 'similar names.' Names that are 'similar' are linked to one another as referring to the same disambiguated institutions.

patent	assignee name	assignee address
EP1807440	Department of Health and Human Services	6011 **Excecutive** Boulevard, Rockville MD 20852
EP2019710	National **Institute** of Health	Office of Technology Transfer 6011 Executive Boulevard, Suite 325, Rockville MD 20852-3804
EP1361886	The Gov. of USA, as represented by the Secretary, **Dept.** of Health and Human services, National Institutes of Health	Office of Technology Transfer, 6011 Executive Boulevard, Suite 325, Rockville, MD 20852

Table 10: Examples of three patent assignees that are "similar" but difficult to disambiguate. A few misspellings or abbreviations are bold-faced. Each of them are geolocated to at least one identical high-resolution latitude/longitudes $(39.048843, -77.120419$ with a quality rating of 87, specifically). There is significant overlap between the words found in the assignee name in EP1361886 and the words in the assignees of either EP2019710 and EP1807440, suggesting these names should be matched.

In order to determine if two names are similar, it is useful to build a list of 'common' and 'rare' words. A dictionary of 'common' words (e.g. "hospital" and "institute") that was generated by hand (and passed through google translate in a few languages) is read into memory, as well as a dictionary of location names (e.g. "Boston") provided by GeoNames [35] and all first and last names occurring in the inventors found in our databases. All words on any of these lists are treated as 'common' in the disambiguation algorithm. A few words in the common list (e.g. "Company") were manually selected as completely uninformative, where the inclusion of these words caused many spurious incorrect matchings, and are removed from the names before matching. For the manually curated common words, a misspelling dictionary is also constructed by (a) producing all deletion errors possible by deleting each character in turn, and (b) by producing every permutation error possible by

swapping the order of every character in the name. These permutations and deletions are *not* applied to the common words. In order to check to see if a name is 'common' within one error, we check to see if it is found in the common word list, or if it *or any single deletion substring* are found in the misspelling dictionary. During assignee disambiguation, words that are on the common list or within one deletion from the misspelling dictionary are treated as 'common,' otherwise they are treated as 'rare.' All of the assignee names are processed as described in Sec. A.7.

To perform the first step of disambiguation, we search for similar names at each high-resolution geolocation. In this draft, "high-resolution" refers to any geolocation with a YQL quality of at least 70, generally corresponding to a geolocation with street-line level accuracy or above. Geolocations with an accuracy on the level of zip-codes, cities, or higher are ignored in this first step of the disambiguation process. The algorithm then iterates over each high-resolution lat/long, and all pairs of names at that location are compared to see whether they are similar:

1. If any two 'rare' words are within one edit distance of one another in both names, the names are a mach. If all rare words in the first name are more than one edit distance from all words in the second name, proceed to step 2. For example, "Harvard University" and "Harvard College" are matched in this step because "harvard" is a rare word.

2. If the list of 'common' words in both names have at least two elements in common, the names are a match (if one of the names consists of exactly one word [e.g. "Apple"] and that word is 'common', only one shared 'common' word is required). If there is not a sufficient overlap between 'common' words found in the names, proceed to step 3. For example, "The General Hospital Corporation" and "Massachusetts General Hospital" are matched in this step, because 'general' and 'hospital' are common words.

3. Check to see if the first name contains an acronym found in the second name. For each word in the first name, break it into individual letters and see if there is a subset of words in the second name (preserving the ordering) that all begin with those letters. If no match is found, check for acronyms in the second name. If no acronym is found at this step, we assume the names are distinct. For example, "The Massachusetts Institute of Technology" and "MIT" are matched in this step, since 'm', 'i', and 't' are found sequentially in the first letters of words in the first name.

Having linked all of the names at each lat/long, every name/geolocation pair is assigned a unique identifier, with any names matched in the algorithm assigned the same identifier. Beginning with the ∼351K name/high-resolution geolocation pairs found in the raw data after geolocation and name cleaning, there are ∼331K unique identifiers produced using this algorithm. We note that this is a rather modest reduction in the number of

name/geolocation pairs, and may appear to have done very little. However, this step performs the essential service of cleaning very noisy names, of particular importance for large institutions. These are the clustered names that are passed to the neighborhood search described in Sec. A.4.

A.3 Inventor name matching

Inventor names generally have less variability in their structure than assignee names, where usually there is a 'last name' (typically the first word in the name), a 'first name' (typically the second word in the name) and finally various 'middle names.' In reality, the first and middle names are sometimes interchangeable, a two-word last name may be separated, and additional titles or company names may be added to the name (see Table A.3 for two examples). In order to overcome these types of errors, we adapt our word-based matching of names at the same high-resolution geolocation that was used to disambiguate assignees. A manual inspection shows that these errors tend to be far more common in the EPO and PCT than in the USPTO (although it is difficult to quantify the rate of any particular error type without an accurate disambiguation in hand), but the USPTO tends to have far lower accuracy in the geolocations as well.

patent	inventor name	inventor address
EP2340782	GOMES DA CUNHA PONCIANO, José Antônio	Av. Ipiranga 55 Centro,CEP: 25685-250 Petrópolis, RJ
EP2386338	GOMES, José Antonio, da Cunha, Ponciano	Av. Ipiranga 55, Centro, Petrópolis - RJ, Cep: 25685-250
EP1247533	Howard, Jr, Harry Ralph	Pfizer Golbal Res. and Dev., Eastern Point Road, Groton, Connecticut 06340
EP1220831	HOWARD, Harry Ralph, Jr. Pfizer Global Reasearch	and Development Eastern Point Road Groton, CT 06340

Table 11: Example of inventor names with a variety of errors. The first two names refer to the same person, but the person's last name is split in the second occurrence (note also the missing accent in Antonio). In the second example, 'Jr' is put in the position of the first name in one instance and a portion of the address field is added to the inventor's name in another instance. Both of these are geolocated to the same position, and a flexible matching can be performed.

In order to disambiguate the inventor names, each name at a high-res geolocation is processed as described in Sec. A.7. A search is performed for the strings "c/o" and "c/-"; any words following this substring are forbidden from being matched at that geolocation under the assumption it refers to the assignee (matching due to that string is not forbidden at other geolocations). In each name, it is assumed that the first two words in the name

correspond to the last and first names, respectively, and we check to see if these assumed first and last names are found *anywhere* in the name we are comparing.

1. Check to see if the 'last name' is found in the name we are comparing to. If the word has more than three characters, allow for a difference of one edit. We keep track of how often the 'last name' was a match to the first word in the compared name.

2. If the 'last name' was found in step 1, check to see if the 'first name' is found in the compared name. If there was *perfect* agreement in the last name, permit the 'first name' match to differ by one edit distance. If there was *not perfect* agreement in the 'last name,' require perfect agreement with the 'first name' (max. edit distance of 0).

For each pair of names, we perform this check using the first two words of each in our search. If we find a 'last name' + 'first name' match in between either of the names being compared at the same high-res geolocation, we link the names as referring to the same individual.

This algorithm is fairly robust and able to disambiguate the names of a majority of high res geolocations without difficulty. However, significant errors can occur due to the confluence of two events in the data: (1) a large number of inventors using the same address and (2) culturally common 'middle names.' For example, the address "Prof. Holstlaan 6,NL-5656 AA Eindhoven" is used as an inventor address for over 34k EPO patents, and at that precise address there are 743 unique inventor names that have a middle name "Maria." In such cases, last names that are one edit distance away from common last names will cause an overwhelming number of incorrect links between names (e.g. there is a person named "Marra, Johannes" at the same address, with a last name one edit distance from "Maria"). To prevent these huge errors, we perform a pruning step, and unlink pairs of names if the 'last name' that caused the link was matched to a non-'last name' more than twice as often as it was matched to a 'last name.' This final post-processing step removes linked names when the 'last name' is overwhelmingly matched to a 'first' or 'middle name', indicating a spurious match.

As was the case in the disambiguation of assignee names, in our final step we assign each name/geolocation pair a unique ID, ensuring that each matched name has the same ID. From the 2,241,414 name/high-res geolocations found in the data, we produce 1,997,388 unique IDs after this round of disambiguation. These identifiers are passed to the neighborhood search described in Sec. A.4.

A.4 Nearby exact name matches

The disambiguations in Secs. A.2 and A.3 provide a robust matching between similar names, but do so only at precise, high resolution geolocations. Many addresses are low

resolution, where e.g. the address "Boston MA" is very imprecise, and would not be used in the name disambiguation described in Secs. A.2 and A.3. Any typographical errors in the address field that could also move precise addresses by even a few meters would prevent the linking of names as well (e.g. "123 Main St." vs "132 Main St."). In order to expand the coverage of the disambiguation to imprecise addresses, we perform a search for nearby geolocations (within 20km of one another) that have an exact name match (after cleaning capitalization and punctuation variations as described in SecA.7). This provides links between high resolution geolocations where identical names are found nearby, links between high and low resolution geolocations of the same name, and low resolution IDs of a single exact name match. This is some sense the analog of the disambiguation described in Secs. A.2 and A.3, which used exact locations to search for similar names; in this step, we use exact names to search for nearby locations.

The names and geolocations (both high quality and low quality) for inventors and assignees are read into memory and processed as described in Sec. A.7. For each unique name, only the highest quality geolocations are kept (e.g. for a name with two geolocations found using YQL, a location of quality 60 and another of quality 39, only the first would be kept). To disambiguate the names:

1. For each name occurring on any patent where YQL provided more than one geolocation of the same quality, we link all of those name/location pairs into the same identity. This corrects any noise in the geolocation on an inventor's or assignee's address on a patent.

2. For each exact name, we determine the distance between all pairs of geolocations. If that distance is less than 20km, link those name/locations.

3. We assign all linked names to a unique identifier, and any alternate spellings of those names that were disambiguated following the steps in Secs. A.2 and A.3 are *also* assigned to the same unique identifier.

4. Finally, we assign any unlinked names their own low-resolution unique identifier.

Many names are never found in conjunction with a high-precision geolocation, and misspellings or alternate spellings in these names will not be corrected using this algorithm. For example, the patents US6495146 and US6028086 have assignee names "Pfizer Incorporated" and "P Pfizer Inc" respectively, both with the address "New York, NY." These names will *not* be linked using this algorithm, due to the simultaneous imprecise naming *and* imprecise addresses. Despite this limitation, we still have good coverage both in the number of disambiguated names and the number of disambiguated individuals and the number of patents they cover, as shown in Table 2.

Linking all identical names at a low-resolution geolocation implicitly assumes that any identical name in the same region (city or zip code) must refer to the same individual.

While this is likely true in many cases, there are also many situations where it may not be true for inventors. For example, two inventors in the vicinity of Warwick, England, are named "Jones, Christopher B" and "Jones, Christopher J." Patents invented by "Jones, Christopher" in this region *cannot* be assumed to refer to the same person, as each patent could potentially be invented by "Christopher B" or "Christopher J". We therefore search for exact name matches for in prefixes of longer names (e.g. "Jones, Christopher" and "Jones, Christopher B"), to identify potential matches due to an excluded middle name as well as potential errors due to linkages in ambiguous cases. To check for prefix matching for both assignees and inventors, we:

1. For each name n in the patent data, we search for exact prefix matches of other names $\{m_i\}$, after removing punctuation and cleaning spacing. In the case of assignee names, we remove non-informative words such as "corp" and "gmbh" in this step.
2. All pairs of IDs that are associated with the name n and one of the m_i's are treated as potentially linked. IDs that are potentially linked to *exactly one* other ID by prefix matching are merged into a single ID.
3. Steps 1-2 are repeated until no new merges occur.
4. After Step 3 is completed, any inventor IDs that are potentially linked to *more than one* other ID are treated as erroneous. These IDs are split apart, with every patent held by the original ID being given its own unique ID. This step is not performed for assignees, only inventors.

The iterative process of linking names with single prefix matches assumes that inventor names may sometimes drop the middle name. The splitting ensures that if two middle names are encountered nearby for the same first and last name, we relax that assumption. The prefix search is also performed for assignees (e.g. "Sony" and "Sony Electronics" would be linked if there was only a single match nearby), but unlike inventor IDs we do not break apart the assignee IDs associated with shorter names if more than one match is seen. This is due to the significant variability in the names of assignees: an ID with the name "Sony" may be close to other IDs with names "Sony Electronics," "Sony Entertainment," "Sony Games," etc. While we do not merge assignee IDs if more than one prefix match is found, it would be unreasonable to break apart the ID associated with "Sony" due to this variability. Testing shows that the quality of the assignee disambiguation is significantly reduced by splitting assignees in this manner, but that splitting inventors does improve all quality indicators.

A.5 Linking Patents without Geolocations to the Disambiguation

The disambiguation methods described in Secs. A.2-A.4 cannot be applied to inventors or assignees that are geolocated with very poor quality or that did not provide an address.

Our geolocation of all addresses was performed on the January 2014 data provided by the OECD, but we have applied this algorithm to the more recent July 2014 OECD database for the EPO and PCT. New addresses would not be included in the geolocation process, and would be treated as being unlocated as well. Incomplete geolocations are generally due to poor quality assignee addresses in the USPTO (see Table A.5).

	EPO	PCT	USPTO
Patents with at least one unlocated assignee	0.11	0.14	1.66
Patents with at least one unlocated inventors	0.29	0.28	0.33

Table 12: Number of incompletely geolocated patents in each office, counts in millions. The overwhelming majority of incomplete geolocations are due to assignees in the USPTO, where address information is often on the city level at best and completely blank at worst.

Each patent having an unlocated name (either inventor or assignee) may still have some inventor geolocations, as some inventors may have provided a high quality address. In order to link these unlocated names to the previous steps of the disambiguation method, we first search for exact name matches between unlocated assignees or inventors, having name n and previously disambiguated assignee or inventor names, and we use the inventor geolocations as a proxy for the unknown name's location:

1. For each unlocated name n on patent p, having a list of $\{x_i\}$ previously geolocated inventors, search for identical names n in the previous disambiguation (after removing capitalization, spelling errors, and non-informative words as in Sec. A.7). Create a potential link from the patent to any ID $\{y_i\}$ having that name n.

2. If any geolocated inventor x_i on this patent is within 20km of a matched ID $\{y_i\}$, link that patent to the nearby ID. If more than one ID is within 20km, link the patent to the one with the most disambiguated patents.

This process links names based on physical proximity, but does not permit any spelling errors or name variations due to the requirement of identical names.

In order to account for noise in assignee (or inventor) names, we also search for disambiguated inventors on patent p that are also inventors on patents with assignees (or inventors) of a similar name. Similarity for inventors in this process is more restrictive than it was in Sec. A.3: we require all words in the shorter name to be found *or* an abbreviation to be found (in both cases words can be in any order, up to one spelling error total in words of length ≥ 4). Assignee similarity is the same as in Sec. A.2. These shared inventorships are used to draw links to the previous disambiguation:

1. Construct a list of patents p that are missing an assignee (or inventor) location and

all disambiguated inventor IDs $\{x_i(p)\}$.

2. For each x_i, construct a list of disambiguated co-assignees (or co-inventors), $\{y_j\}_{x_i(p)}$.
3. For each patent missing a location on name n, search for the most commonly occurring ID y_j that is similar to the unlocated name n (using the methods of Sec. A.2 for assignees and the more restrictive name matching for inventors).

This uses our disambiguation, rather than inventor geolocations, to attach an unlocated name on a patent to a disambiguated ID. This step permits some spelling correction of names when linking the patents to the disambiguation, but as there is no geolocation information on the linked patents, we do not treat the linked IDs as 'high-quality' in our analysis.

Note that this second procedure can introduce noise in the number of raw names that are associated with an ID. For example, "The United States of America as represented by the Department of Health" and "The United States of America as represented by the Department of Veterans Affairs" are by all respects similar names: the edit distance between them is 80% of the shortest length, and the number of identical words is over 80% of the total number of words in them. An inventor that predominantly works with the Dept. of Health and Human Services (HHS) but who collaborates with the Dept. of Veterans Affairs may cause these two entities to be linked incorrectly. For the HHS, we find these sorts of errors account for well under 5% of the total number of patents assigned to its disambiguated ID, so these errors are relatively rare.

A.6 Mobile Inventor Disambiguation

The approaches to disambiguation in Secs A.2-A.5 provide a disambiguation of institutions and inventors, but only within a ∼20km radius. While assignees are generally expected to remain in a fixed and localized position (neglecting the possibility that the same company moves from one city to another), inventors are far more likely to move between cities or countries. Our algorithm would be incapable of linking the names of mobile inventors, since the distance between their geolocations would be well above our threshold of 20km. In order to overcome this limitation, we must add one final step to the disambiguation algorithm that allows for inventor name matching over greater distances. We do this by searching for similar names that have other important characteristics in common, suggesting that they are the same person. In particular, we link a pair of disambiguated inventor IDs that share an *exact* name match if any of the patents held by those two inventors

1. share a disambiguated co-inventor
2. share a disambiguated co-assignee

3. are members of the same triadic family
4. have at least one citation from one to the other

Note that shared technological classes are not included in our measure of similarity, due to the varying levels of aggregation available (raging from a handful of WIPO industrial sectors to tens of thousands of 7-digit IPC codes) and the unknown technological and cognitive distance between classes.

Inventor names that are similar (requiring all words in the shorter name to be found in the longer name) are taken to be the same individual if they have *two* of the characteristics in common. While the primary purpose of this step is to detect the long-range mobility of inventors between regions or countries, we note that this permits spelling error correction in inventor names within the 20km range that were not possible in step 2 of the algorithm. IDs linked in this step are thus labeled as 'high quality,' as they incorporate some level of geolocation and spelling correction.

A.7 String Handling

When parsing the names, all latin characters are converted into lowercase letters. All accented letters are converted into the character 'x' (this has the affect of penalizing the removal of an accent, but not penalizing the change of an accent in a name). All symbols (e.g. '-' or '.') are converted into spaces. After this, words composed of a single character are deleted and all double-spaces, triple-spaces, etc. are converted to a space.

For assignees, a specific list of $\sim$ 60 words are deleted from the names before passing through to disambiguation (for example, "inc", "corporation", "aktiengesellschaft", and "the"). The designation of words to completely delete before disambiguation were made manually, and altering this list may change the results to some extent.

For inventor names, a few words are dropped from the names: "dr", "de", "da", "di" "mc", "von", "der", "van", and "den". These short words tend to be non-informative within the last name of individuals, and cause problems with our algorithm if they are treated as a 'first name.'