RIETI Discussion Paper Series 18-E-018

Inventor Name Disambiguation with Gradient Boosting Decision Tree and Inventor Mobility in China (1985-2016)

YIN Deyun
University of Tokyo

MOTOHASHI Kazuyuki
RIETI

The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 18-E-018

March 2018

Inventor Name Disambiguation with Gradient Boosting Decision Tree and Inventor Mobility in China (1985-2016)[1]

YIN Deyun (University of Tokyo) and MOTOHASHI Kazuyuki (RIETI and University of Tokyo)

Abstract

This paper presents the first systematic disambiguation result of all Chinese patent inventors in the State Intellectual Property Office of China (SIPO) patent database from 1985 to 2016. We provide a method of constructing high-qualitative training data from lists of rare names and evidence for the reliability of these generated labels when large-scale and representative hand-labeled data are crucial but expensive, prone to error, and even impossible to obtain. We then compare the performances of seven supervised models, i.e., naive Bayes, logistic, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), as well as tree-based methods (random forest, AdaBoost, and gradient boosting decision trees), and found that gradient boosting classifier outperforms all other classifiers with the highest F1-score and stable performance in solving the homonym problem prevailing in Chinese names. In the last step, instead of adopting the more popular hierarchical clustering method, we clustered records with the density-based spatial clustering of applications with noise (DBSCAN) based on the distance matrix predicated by the GBDT classifier. Varying across different testing data and parameters of DBSCAN, our algorithm yielded a F1-score ranging from 93.5%-99.3% with splitting error within the range 0.5%-3% and lumping error between 0.056%-0.37%. Based on our disambiguated result, we provide an overview of Chinese inventors' regional mobility

Keywords: Disambiguation, Patent, Inventor, Machine learning, Gradient boosting, DBSCAN

JEL classification: C60, J61, L30

[1]This work is mainly supported by the Research Institute of Economy, Trade and Industry's (RIETI) under the project "Empirical Analysis of Innovation Ecosystems in Advancement of the Internet of Things (IoT)," NSFC-JSPS Scientific Cooperation Program between China and Japan (No.71711540044) and the National Natural Science Foundation of China (No. 71503123). We also express our appreciation for the helpful comments by the participants at a RIETI discussion paper seminar, a NISTEP seminar, and NSFC-JSPS project meetings, particularly by Kenta Ikeuchi (RIETI).

1. Introduction

After 30 years of rapid growth (over 18% on average), patents applied to the State Intellectual Property Office of China (SIPO) is promised to become the largest patent database in the world within 3~5 years. Although some valuable indicators like citations are absent, this huge dataset holds abundant information for studying technology, innovation, and entrepreneurship in China. Patent data of USPTO (the United States Patent and Trademark Office), EPO (European Patent Office), and JPO (Japan Patent Office) have served as a major source of evidence for empirical studies about knowledge production and spillover through co-inventing networks, inventor mobility, and technological cooperation for a long time. And endeavors on disambiguating patent inventor data which is initiated by Hall (Hall et al., 2001; 2007), Torvik (Torvik & Smalheiser, 2009), Singh (Singh, 2005), Lissoni (Lissoni, et al., 2016) and Fleming (Fleming, 2007; 2009) have pushed the frontier of studies extending from firm-level to inventor-level. In contrast to these fruitful innovation studies, individual-level researches on Chinese inventors remains rare. A major obstacle is the lack of disambiguated inventor data.

Name disambiguation, or entity resolution, is a problem of identifying "who owns which", namely, drawing a boundary for all patents (or papers) owned or participated by the unique person while excluding others that do not belong to. According to Ventura et al., disambiguation, which links records of unique entities (people or organization) within a single dataset (find duplicate entities), can be considered as a subset of the broader "record linkage" problem that link records of unique entities across multiple datasets (Ventura et al., 2015).

Disambiguation is a preliminary and fundamental step of micro-level research on topics involving inventors' productivity, mobility, and collaboration network or linkage with external datasets of papers or surveys. When searching the name "张伟" (Zhang Wei) in the SIPO database, for example, you will receive 9,680 patent records. Obviously, as a common name in China, these patents could not be invented by one person. Without extra information, we do not even know whether two patents under the same name refers to the same person or two distinct ones. Analysis depending on such kind of data would be highly questionable. In addition to

benefiting the academic community[1], disambiguation of inventors' names could also yield more accurate querying results and assist head-hunters in identifying unknown but productive and talented engineers, discovering potential cooperation partners or business opportunities, or assist governments and research institutes in constructing invention profiles of inventors and evaluate organizations and inventors' performances.

Nearly any form of disambiguation would encounter these two types of challenges: the synonym problem, in which the same person appears with several distinct names because of name change, abbreviations, typos, or misspelling, and the homonym problem (the polysemes, or the common name problem), i.e., many distinct people are referred by one same name (Han et al., 2017; Louppe et al., 2016; Müller et al., 2017). As Chinese names barely have middle names and abbreviations and thus less variant, the synonym problem, although very common in names written in English, is relatively trivial[2] in Chinese texts. Thus, considerable efforts of extent disambiguation work were devoted to compare the similarity of first and middle names and finding an appropriate blocking strategy to capture variants of names as much as possible. In Chinese names, the most prevailing and troublesome problem is the homonym problem. The National Population Census in 2000 shows 84.77％of the population has one of the top 100 family names in China, whereas in the United States it is 16.4% (Wikipedia[3], (Kim, Khabsa, & Giles, 2016). Specific to patent inventors, there are 0.61 million records (4.20%) referred by

[1] Disambiguation of inventors would open new research areas and provide micro-level evidences for empirical studies. For example, it can help identify the most productive ("star") engineers (prolific or with higher influence) and examine factors that influence organizational or inventor's productivity, and enable the tracking of inventors' patenting careers and mobility among institutions and regions, thus allowing scholars to examine the determinants of inventor mobility and its impact on knowledge spillover or firm performance. It can offer evidence for knowledge flow through inventor mobility and collaboration network as well.

[2] In the rare name dataset, only 0.2~0.25% record has synonym problem and we suppose this problem in the whole dataset would be fewer than 0.2% as rare names are prone to be misspelled or miswritten and common name constitute a larger portion.

[3] https://en.wikipedia.org/wiki/List_of_common_Chinese_surnames

names corresponding to 500~1000 patents and 0.96 million records (6.56%) referred by names with 500~9680 patents in SIPO dataset, all of which are much higher than that in USPTO and JPO, as summarized in Table 1.

(Table 1)

Although a very small portion of disambiguated Chinese inventors from USPTO and PATSTAT are available[4], the presence of East Asian names in large quantities and those distinct characteristics listed above are acknolowleged as the primary challenge for the US-centric or Europe-centric methodologies (Tang & Walsh, 2010; Wang et al., 2012; Li et al., 2014). Some disambiguation practices hold that East Asian names, especially Chinese names, should be processed separately or at least add the ethnic dimension of inventors to improve the disambiguating accuracy (Chin et al., 2014). Furthermore, when Chinese names written in Chinese characters are translated into Latin characters, they lose their valuable identifying properties, and Type II error (falsely predicate two different persons as the same one) would increase significantly. Accordingly, directly working on original Chinese names is supposed to produce better classification results.

In addition to those common challenges and unique problems caused by the characteristics of Chinese names, Chinese patent dataset also contain no information on inventors' addresses and the citation relation provided by inventors or applicants, which are two important and necessary components in major patent datasets across the world[5] and weighty features to improve disambiguating accuracy (Ferreira, Gonçalves, & Laender, 2012). To fill in the lacked

[4] In USPTO data cleaned and disambiguated by PatentsView, there are 61,465 inventors locating in China. Data checked from http://www.patentsview.org/web/#search&loc=China&loc-type=inventor on 2017.12.31. Considering the data coverage, unknown accuracy and potentially high error rate of existing results about Chinese inventors in databases such as USPTO or PATSTAT, a disambiguation of Chinese inventor names processed according to their own characteristics is still necessary.

[5] Degree of precision of inventors' location information vary among different intellectual property offices: for example, USPTO offers state-level while EPO and JPO have detailed address even including room number or street number. Inventors can choose whether to report their work or home address. Lacking this data would obstruct studies involving tracking geographical information. Concerning the citation data of Chinese patents, SIPO began to provide examiner-added citations for patents granted after 2008. Nonetheless, currently, no inventor- or applicant-added citation is available.

dimension of inventors' addresses, we adopted the location (including geographical coordinates and addresses parsed six administrative levels) of the first applicant in SIPO patent dataset as a substitute.

Finally, in contrast to the disambiguation of papers which has millions of labeled records (Louppe et al., 2016), labeled data of patents is few and precious. While disambiguation work conducted on USPTO dataset to which many scholars contributed up to 5 training and testing data from different industries or technological fields (of course, most of them belong academic inventors)[6], we have no data for training a model and evaluating classification results. Of course, because all of these data are not collected for disambiguation purpose, as Ventura et al. (2015) have argued, models trained and evaluated on these datasets yield varying scores and suffer bias toward some industry or prolific academic inventors. To overcome these shortages and biases and to provide a reliable database for academic and industrial analysis, we selected 66,248 rare names which correspond to 402,339 inventor-patent records, as well as 21,073 inventor-patent records participated by 1,314 academic and industrial inventors from a broad range of technological fields.

In this article, we present 1) the first systematic disambiguation of Chinese inventors with machine learning algorithms customized according to characteristics of Chinese names; 2) a comparison of seven supervised learning classifiers and their performances with different training sets and input features; 3) a discussion about performances of different clustering method (hierarchical clustering vs. DBSCAN) on Chinese names; and 4) evidence for utilizing statistically generated data as substituting for hand-labeled records in training models when large-scale, representative and low-biased hand-labeled records are difficult to collect. Besides, we also analyze Chinese inventors' international and regional mobility based on the disambiguated results.

---

[6] In May ~ September, 2015, PatentsView hold a disambiguation workshop and competition to improve the value and utility of USPTO patent dataset. They gathered these dataset could be downloaded from the following website: http://www.patentsview.org/workshop/participants.html#data

Besides the methodological part, we contribute to the research community the following datasets: 1) A cleaned dataset of Chinese inventors with the ethnic classification (Chinese, Japanese, or other foreign names); 2) The first harmonized applicant name based on the string similarity of applicant names' stems and the geolocation information of the first applicant. We hope this result would benefit works aiming at linking firm financial and other information with SIPO patent database. And 3) a hand-labeled dataset with low-bias towards particular industries for evaluating and comparing performances of disambiguation methodologies.

This paper is organized as follows: Section 2 provides a thorough review of disambiguating works and algorithms in the order of algorithms' type, time of publishing, and datasets scholars worked upon. In section 3, we describe our data and methodology step by step. Section 4 evaluates our algorithm's performance both before and after the clustering stage with F1-score , as well as the splitting error (similar to Type I error, falsely predicting two identical records as different) and lumping error (similar to Type II error, falsely predicting two distinct records as the same)[7] on testing data. We also compare seven kinds of widely used supervised classifiers and their performance with three kinds of training sets. We present results of selecting the most appropriate training data, the best combination of features and model after 5-fold & 10-fold cross-validation. Based on our disambiguated result, we then provide an overview of Chinese patent inventor dataset and some analysis about inventors' mobility in section 5. The last section concludes.

2. Literature Review

Disambiguation of persons' names originates from the demand of digital libraries (DL) in disambiguating authors of papers. Frankly speaking, all works that disambiguating patent inventors are built on the pioneering explorations and abundant methods proposed by researchers in the area of DL. Ferreira et al. (2012) provided a thorough review of methods on author disambiguation before 2012, including heuristic approaches (Pereira et al., 2009), supervised

[7] Please check Section 3.3 for details.

methods like Naïve Bayes probability model and Support Vector Machine (SVM) of Han et al., (2004), Hierarchical Naïve Bayes Mixture Model (Han et al., 2005), SVM-DBSCAN (Huang et al., 2006), Random Forest (Treeratpituk & Giles, 2009), as well as unsupervised methods like K-way spectral clustering model (Giles, Zha, & Han, 2005). However, since there are many differences in demands and type of information available (different format of names, lacking information like inventors' address and citation) between these two tasks, this article mainly focuses on works conducted on patents.

Based on whether an algorithm makes use of labeled records to train its models, methods for disambiguation could be classified as supervised and unsupervised. Simply speaking, when all input data are labeled records (have a known result), the method belongs to supervised learning, and unsupervised learning refers to the situation in which there is no labeled input data. Semi-supervised is something in-between: its input data is a mixture of both labeled and unlabeled records. In unsupervised learning, disambiguation is treated as a clustering issue (Wang et al., 2012), while in supervised learning, it is a binary classification problem in the first step and then records could either be clustered according to some particular threshold or based on affinity or distance matrix predicted by classifier trained with labeled data.

However, before machine learning techniques arose, methodologies adopted, such as that of Singh (2005), Trajtenberg et al., (2006), Fleming et al. (2007), Jones (2009), Cassi & Carayol (2009) and Raffo & Lhuillery (2009), etc., are mainly rule- and threshold-based: i.e., based on expert knowledge, researchers develop a set of rules, add some ad hoc weights and thresholds to determine whether two records should be linked (Ventura et al., 2015). Since most of rule-based methods or heuristic approaches often lack training data, they are usually classified as a kind of unsupervised method. However, they could either be supervised or unsupervised according to whether they make use of training data. For instance, Pezzoni et al.'s (2014) name their method as supervised rule-based method as they used two datasets in training their thresholds. At the same time, a typical disambiguation algorithm is composed of four components as presented in

Figure 1: 1) Blocking, 2) Computing similarity scores (or similarity profiles)[8], 3) Combining scores of many different dimensions into a single distance with or without a trained classifier, and 4) Clustering (or author assignment in (Ferreira et al., 2012). The parsing/cleaning-matching-filtering procedure of rule-based (heuristic) approach summarized by Raffo & Lhuillery (Raffo & Lhuillery, 2009) could find their correspondence to blocking and computing of similarity scores. Because the boundary between rule- or threshold-based approaches with machine learning approaches is fuzzy, we would introduce the former separately and regard it as a prototype or early-stage of machine learning approaches.

(Figure 1)

Table 2 summarizes existing work according to this definition. For more information about their self-reported evaluation scores, please check the Appendix 1 at the end of this article.

(Table 2)

To date, most of disambiguation works and approaches are conducted upon USPTO inventor data. And it is also on USPTO data disambiguation algorithms gain diversity and significant improvements in accuracy, speed, and sophistication. For disambiguation works on international patent databases, like PATSTAT[9], the rule- or threshold- based approach prevails as representative training sets with small biases towards large patent-filing countries like US, Japan, and China is extremely hard to obtain.

2.1 Rule-based approach

Research teams in Bocconi University, represented by Pezzoni & Lissoni, have been dedicated to the cleaning and disambiguation of inventors of PATSTAT for more than a decade. Since their

[8] Blocking is a step that widely adopted for computation-intensive works. It is not an indispensable component. Disambiguation methods could differ from each other at each step. Rule- and threshold-based approach in the early-stage usually lacks the blocking or similarity profile building step.

[9] Worldwide Patent Statistical Database (PATSTAT) is the major data source for European and global patents which is released by EPO periodically.

dataset construction and disambiguation work made use of different raw data sources, it went through many names from “EP-INV” to “APE-INV” (Pezzoni et al., 2014). Their algorithm for disambiguating inventors, named as “Massacrator”, evolves from purely unsupervised rule-based method in the beginning (Lissoni et al. 2006) to supervised rule-based method which finds the threshold of being match (1, positive case) or non-match (0, negative case) by comparing its results with two benchmark datasets of academic inventors. Although their precision rate between 56%~92% and a recall rate between 93%~54% seems pretty bad compared works that claim their precision and recall are all above 90% or even 98%, on the one hand, it reflects the disadvantage of rule-and threshold-based approaches to machine-learning techniques. On the other hand, it is quite understandable as PATSTAT patent database covers patent records from many countries with various languages, it is remarkably difficult to capture all name variants by a rule or collect a representative training set to lower the splitting error across different languages.

Pezzoni et al., (2014) also indicate that thresholds chosen based on inventors’ addresses are sensitive to data quality while those based on co-inventor networks yield more robust results. Morrison et al. thus overcame this problem and proposed a simple and unified rule to disambiguate 8.5 million patents of EPO, patents under the Patent Cooperation Treaty (PCT) and USPTO based on high-resolution geographical information (Morrison, Riccaboni, & Pammolli, 2017). Although they made use of other patent information along with inventor names and addresses, this approach made strong assumptions and relies heavily on the role of inventor address in determining linkage results.

Among the fruitful works adopting rule-based methods, it is Tang & Walsh (2010) who began to give special attention to the homonym problem predominant in Chinese names. They built an “ASE” algorithm (approximately structural equivalent) that disambiguates inventors based on the citation relation (“bibliometric fingerprints”). Han et al., (2017) improved this method and disambiguated Chinese authors with semantic fingerprints (keyword metadata or research topics) extracted from text, co-authors and institutional information.

In the case of Chinese inventors, Gupeng Zhang and his team made a rough disambiguation of SIPO inventor data between 2000 and 2009 in their study of inventor network in China

(Zhang, Guan, & Liu, 2014). Instead of selecting Chinese inventors by name, they filtered out 0.188 million patents with addresses in China and provided a detailed description of their rules. Their rules are as follows: to begin with, they divide patent data into 330 cohorts according to 10 applications a years and 33 provinces where the first applicant locates in, and then identify inventors within these 330 cohorts. For inventors with the same name, if they also have the same job and address, they are identified as the same person. For those who do not satisfy this requirement, check whether they have a similar technological field or common co-inventors. After these steps, 1% names with only two patents remain. They identified 0.89% inventors by connecting the firms that these inventors serve to check their work experience via phone call. The rest 0.11% are omitted from the database as the ratio is relatively small and would not have large impacts on their empirical result.(Zhang et al., 2014)

Gupeng Zhang's team provided the first and unique attempt to disambiguate Chinese inventors with SIPO data. Unfortunately, they did not provide their data or evaluation of their method. Many technical details need discussion[10]. However, as disambiguation is not their focus but just a way of data construction, no further critiques should be imposed on their achievement. In addition to the common problem of the rule-based method and its large lumping error, the sheer bulk of patent data (4.9 million patents with 14.68 million inventor-patent records during 1985-2016) which increases with incredible speed has rendered this approach inefficient and infeasible.

In summary, the rule-based method is simple, intuitive and lower in computational burden while it usually relies on expert knowledge and simplifies complex situations by a set of rules. These rules often contain strong assumptions or assign the decisive rule to some particular

[10] What need to be discussed about Zhang et al.' method are: first, they did not mention how they solve the transitivity problem of inventors across different years and provinces as it is natural for inventors to move across different places and change affiliations. Second, there could be hundreds of thousand people share a common name within a province in China. Consequently, if they adopt province as the address there would be very high lumping errors. Third, still, their phone call visit of inventors' firms is not direct investigation of patent participation or ownership. Human judgment is still necessary for their last step.

features. However, there are neither perfect rules nor such determinant features. There are always exceptions that could not be captured by the combination of rules. Such assumptions, which are easily violated in the real world, lead to systematic errors and bias unevaluated in results. This explains partly why rule-based methods are usually beaten and gradually substituted by machine learning approaches, for which neither too much domain knowledge nor rules with strong assumptions are required.

2.2 Semi-supervised learning

After trying the rule-and threshold-approach in 2006 and 2009, Lee Fleming and his team members proposed the first semi-supervised algorithms on inventors with a measurement of its performance in 2014 (Li et al., 2014). Their pioneering work in 2014 followed Torvik et al.'s Author-ity approach (Torvik et al., 2005; Torvik & Smalheiser, 2009), which selects a list of rare names according to the number of patents that each full name corresponds to and constructs matching and non-matching pairs based on this list. Mixed with this statistically generated artificially labeled records, they applied their Naïve Bayes classier as well as some ad hoc rules to disambiguate 9 million inventor-patent records. Naïve Bayes classifier is very suitable for big datasets like patent databases, and it usually outperforms other approaches by its speed, simplicity, and accuracy. They also evaluate this method with 1,169 inventor-patent records filed by 95 eminent academic inventors from engineering and biochemistry fields.

By opening their data and code to the public, they not only provide a large-scale and wide-covering data to support inventor-related empirical studies but also inspire more scholars to apply or refine this methodology into different datasets. Ikeuchi et al. (2017), for example, applied their customized algorithm of Li et al. (2014) to all Japanese inventors in JPO.

Semi-supervised learning from these statistically generated labeled records would usually receive more information near the decision boundary for classification and thus yield better results compared to a set of decision-rules or unsupervised learning. However, its Achilles' heel also lies in these statistically generated labeled records: first of all, the selected rare names are only rare within patent data, while that may not be the case in the real world. For the SIPO database, we cannot say that those 0.7 million unique names that correspond to only one patent record (or the 0.15 million unique names corresponding to two records) are rare or rarer

compared to names corresponding to several hundreds of records. Second, the "highly likely" to be matches and non-matches are built according to their similarity of assignees and addresses. However, perhaps this clear boundary delineated with rules about two or three features could not capture the non-linear boundary between two classes. In a word, labels generated from these rules would either contain many mistakes due to exceptions to rules or being too simple to be classified. Therefore, these mistakes would either be reflected and amplified into or provide little information for disambiguation results. Luckily, Ikeuchi et al. (2017) provide a promising solution on these statistically generated labels by extracting rare names from telephone directory during 2000-2012.

2.3 Supervised learning

As more scholars delve into studies about inventors, their manually collect data would contribute valuable data for training and evaluating disambiguation methods. Based on 98,762 labeled records of industrial inventors in optoelectronics (OE, provide by Akinsanmi et al. (2011)) and 53,378 records of academic life scientists (ALS, provided by Azoulay et al. (2011)), Ventura et al. proposed the first supervised learning method with the random forest classifier and single linkage hierarchical clustering algorithm on USPTO dataset (Ventura et al., 2015). Their comparisons with other approaches illustrate that random forest classifier yield the lowest errors consistently among multiple supervised classifiers and their supervised algorithm outperformed the rule- and threshold-based approach adopted by Li et al. in 2007 and 2009 (Fleming et al., 2007; Lai & Fleming, 2011) as well as the semi-supervised approach proposed by Li et al. in 2014 on subsamples (not the full)[11] of OE dataset.

Ventura et al. also decomposed their results by training and testing their model of purely industrial inventors (OE), purely academic inventors (ALS), and a mix of them (OE+ALS), and discussed sources of bias in previous algorithms. Their comparison results manifest that dataset composed purely of academic inventors could neither serve as a qualified training set as it carries

[11] In fact, Li et al.'s (Li et al., 2014) semi-supervised approach out-performed Ventura et al's, (2015) supervised approach slightly on the full data of OE and ALS.

insufficient information and thus have higher error rates on unknown data, nor serve as qualified testing set because it's much easier to be classified and the real error rate based it would be underestimated. However, it is also these defects of their data prevented them from applying their algorithm trained on the OE data on the full USPTO data: these two datasets are not collected for disambiguation (so they missed patents belonging to the same inventor but filed outside of these two industries) and they have large bias towards one particular industry or prolific academic inventors. Otherwise, new defects and bias in this training set would be further introduced into disambiguation results.

On September 2015, PatentsView held a disambiguation competition among six teams[12]. Nicholas Monath and Andrew McCallum's "Discriminative Hierarchical Coreference" algorithm won the contest, and now their method has been incorporated into PatentsView's disambiguation algorithm and results (Monath & McCallum, 2016).

Before Ventura et al. (Ventura et al., 2015), Treeratpituk & Giles(Treeratpituk & Giles, 2009) already demonstrated how random forest outperform other techniques such as SVM. They disambiguated 4 million paper authors (12 million paper-author records) within 24 hours with the random forest as a pairwise classifier and used DBSCAN to cluster records (Khabsa, Treeratpituk, & Giles, 2014). They then improved this methodology and applied it to 1.2 million USPTO inventor records, the whole process of which was conducted within 6.5 hours (Kim et al., 2016). With the same training and testing sets provided by PatentsView's workshop, Kim's team claim they achieved a better pairwise F1score (98.82% on average) over five testing sets than that of Monath and McCallum (98.27% on average). Moreover, teams participating in PatentsView's workshop also report a running time of 7 hours for the entire process[13]. According to papers and other public information available, the time spent on disambiguation of the main

[12] http://www.patentsview.org/community/workshop-2015

[13] These data is received from http://slideplayer.com/slide/8193569/ .The running time ranges according to the amount of data being processed, configuration of computer (memory, CPU and speed of hard disk) and algorism (language and quality of codes, blocking strategy, number of features involved, etc.)

patent database (USPTO, PATSTAT, JPO, and SIPO) with machine learning approach ranges from 3.5 to 10 hours on consumer-level PC or entry-level workstation (Kim et al., 2016). In other words, the advantage of rule-based approaches over other algorithms in terms of speed is shrinking and disappearing.

In addition to random forest, scholars also tried other ensembled tree methods like AdaBoost and Gradient Tree Boosting (Wang et al., 2012) in author disambiguation and receive excellent results. Louppe et al., (2016), for instance, applied Gradient Boosting classifier and hierarchical linkage clustering to disambiguate authors of high-energy physics[14].

To wrap up, existing studies demonstrate that supervised methods achieve better results over other approaches when high-quality training set is available. (Ferreira et al., 2012; Han et al., 2017) Conversely, if the training data has large biases, these biases would also be propagated in results as Ventura et al. (Ventura et al., 2015) had criticized on other approaches. The quality of training samples has become the bottleneck of supervised learning algorithms. Under the circumstance that large-scale hand-labeled data is costly to collect and contains sizable bias, could the rare name data provide a promising solution as a training set? Which datasets should be used when both of them have defects?

2.4 Unsupervised learning

As shown Figure1, the last step of disambiguation is grouping records either based on predefined similarity functions or learned from a classifier. There are four kinds of extensively used clustering algorithms: 1) Partitioning: which cluster records based on the pre-specified number of clusters, for example, K-Means clustering technique. 2) Hierarchical clustering, which clusters records iteratively in a hierarchical way. 3) Density-based clustering, for instance, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Huang et al., 2006; Kim et al., 2016) and 4) Spectral clustering (Giles et al., 2005).

[14] To be more accurate, disambiguation algorithm could be either supervised, semi-supervised or unsupervised at both the classification and clustering stage, as manifested in figure 1. Louppe et al.'s, (2016) method is semi-supervised in the clustering step.

After Li et al. (2014), Lee Fleming's team members also tried the unsupervised K-Means clustering algorithm in inventor disambiguation and automated the processing (including cleaning, parsing and disambiguation of assignee and inventors) of all weekly updated USPTO data (Balsmeier et al., 2015). However, except that of (Huang et al., 2006; Kim et al., 2016), the majority of disambiguation work adopt the single linkage hierarchical clustering to cluster inventors based on distance matrix, as hierarchical clustering does require pre-specified number of clusters and it could also resolve the transitivity of pairwise matches. By recasting transitivity as density reachability in DBSCAN, Huang et al., (2006) suggest that DBSCAN could solve this problem effectively.

3. Methodology

3.1 Data

3.1.1 SIPO patent data

The data we work on is the invention patent dataset offered by SIPO from 1985 to Dec.31, 2016. It contains the application number, names, and addresses of applicants, inventors, IPC code, title, abstract, claims, lawyer and other information of 6.25 million unique patents. After laborious preprocessing, such as removing extra spaces, strange symbols, and samples generated during recognizing from original patent files, we integrate these separate datasets into a huge one composed of 18.66 million inventor-patent records. Here we set the patent-inventor pairs (the patent authorship, or inventorship) as the unit of data and name it as "ida_seq" (generated by combining application id with inventor's sequence in this patent) for short.

From the SIPO inventor dataset, we then filter out 1.8 million Chinese names (including inventors from some Asian countries like Korea and Japan whose names cannot be differentiated with Chinese names[15], as well as ethnical minorities in China). Chinese names are distinguished

[15] Korean and Japanese names could be identified from Chinese ones easily if they are written in English. However, for names written in Chinese, distinguish them simply by names' properties is formidable if no extra information is available.

from those of Japanese & other countries names by the characteristics of names such as whether there are points inside names, length of characters, their family names and so on. Among the 14.675 million Chinese inventor-patent records, only 13.97 million must be disambiguated as there are 0.7 million names referring to only one patent. Table 3 summarize our classification rule and statistics of data. It also reveals that the common name problem is most severe among Chinese names and thus an ethnicity-sensitive name processing is necessary for disambiguation records involving enormous Chinese names.

(Table 3)

3.1.2 The hand-labeled data

To calibrate our algorithm, we manually labeled patents of 1,314 inventors in the following way: first, we search for inventors reporting patents they owned or participated in by keywords “专利” (patent) and then filter out those who really list the exact patent number by searching keywords “CN”, “ZL” (part of application or publication number) or “专利号” (patent id) from CV or online profiles of academic inventors posted on university websites, and resume of job seekers from trustworthy online recruiting platforms like Liepin.com and LinkedIn. According to information like patent number, inventors’ working careers and news about the inventor and his/her inventions, we then label patents with their corresponding inventor’s name in SIPO. Based on these web pages we labeled 21,073 inventor-patent records with 929 unique names covering both academic and industrial inventors under the principle of selecting samples as random and diversified as possible. During and after the data-collection, 5 rounds of manual checks based on original self-reported information, patent data and similarity scores with other records were executed to ensure accuracy. The lesson we learned from these repeated check and correction of hand-annotated data is that manual labeling is tedious, prone to mistakes, full of uncertainty (because it highly depends on the availability of inventor’s information on the internet) and sometimes outplayed by computers when the information is too much to be

processed by the human brain. Unavoidably, we added our judgments into this collection process when self-reporting data are dubious, incomplete or fail to be updated[16].

Limited by the availability of public self-reported data on patent inventorship, especially those from industries, we obtain a dataset composed of more academic inventors than industrial inventors and the majority of academic inventors concentrated within Zhejiang University, Shanghai Jiao Tong University, or being a fellow of the Chinese Academy of Science (CAS) or Chinese Academy of Engineering (CAE) which signifies the status of academic elite in China. Inventors from some second- or third-tier universities like Northwestern Polytechnic University were also included, albeit their numbers are relatively smaller. Records under industrial inventors cover firms with a wide range of size, location, and industries. Although they only constitute 23.68% of hand-labeled inventorship records, this part of data is more diversified and representative. Interestingly, in contrast to Ventura et al.'s (2015) discovery, China's academic inventors tend to be three times prolific than inventors from industry.

Comparison pairs in which two inventor-patent records have the same id and label are defined as "match" and those who do not are "non-match". To have enough "non-match" instances, we fetched all samples sharing the same name with records in our hand-labeled data and then the sample size rise to 128,753 inventor-patent records.

This hand-labeled data avoids the problem in existing works mentioned above: it was collected for the aim of disambiguation, has little bias towards one particular industry, covers both academic and industrial inventors, contains records with name changes and typos, and it is similar to the unlabeled records remained in entire SIPO Chinese inventor data. Admittedly, these data are not 100% accurate as we did not connect with inventors in person and added our

[16] To illustrate, when two records have the same name, we would judge whether they refer to the same person based on inventor's self-introduction, working career, self-reported patents and patents' information in SIPO like whether they have the same applicant, number of co-inventors, title and abstract of patents. When we are not certain about the result, we would continue to search for other public information like news to find extra evidence. For those judgement made with under 95% confidence, we would simply give up labeling this person.

judgment for those confusing instances. In addition, its amount is limited compared to the size of the vast majority of unlabeled ones. The data constitutes only 0.92% of all 13.96 million records that need to be disambiguated.

3.1.3 The rare name data

Following Ikeuchi et al.'s (2017) suggestion, this paper also constructed a dataset with much larger amount and lower bias automatically: the rare name data. We constructed this dataset by combining Chinese inventors' name list of SIPO with a list of rare names fetched from "same name" websites which provide the information about the counts of people with a particular same name in China[17]. We also made use of other public online information related and removed Korean names mixed in this list manually. To filter out a precise list of rare names, we impose a rather strong criterion: only names belonging to Chinese names and corresponding to at most one person in China would be accepted as rare.

Even though the rare name data is more representative and larger in quantity, it has some differences with the remaining SIPO inventor data. First of all, while it is easy to construct positive or matching pairs from this rare name data since all records under the same or very similar rare names belong to the same unique person, negative or non-matching pairs are insufficient. Here we build non-matching pairs by comparing one rare name with other different names. Accordingly, the rare name data is different from the remaining unlabeled records which would be compared within the block of the same or very similar names. Furthermore, these artificially constructed non-matches would have lower similarity score and thus easily identified as non-matches. To avoid our models just learn from this "easy" distinction, we added one to the number of common inventors for all non-matches. As a result of such limitations existing within the rare name data, we would only use it as the training data.

---

[17] We select the list of rare names by combining information about the number of people one name corresponding to in China from websites like http://www.sosuo.name/tong/ , public data offered by Guozhengtong on (http://zhaoren.idtag.cn/samename/searchName!searchIndex.htm) and free services provided by province and city-level public security bureau. Unfortunately, Guozhengtong had stopped the public access to this website in 2017.

Table 4 displays the repressiveness of these two datasets in terms of applicants' type and Table 5 provide a brief description of them, including the large SIPO Chinse inventor database. From these two tables, it is not difficult to figure out that these two datasets are superior to many other datasets discussed in the last section in quantity and low-biasness, although they also suffer from some shortcomings.

(Table 4) & (Table 5)

3.2 Disambiguation algorithm

Building on experiences of these efforts above, we devise our algorithm with the following four steps:

(Figure 2)

3.2.1 Blocking

A full comparison of each of 13.96 million records with all other patents means that we must compute the similarity of 98 trillion pairs, which would impose extremely high and unnecessary computational burden. Blocking, which means dividing inventor-patent records into disjoint subsets according to some rules and conducting comparisons within each block, is a popular strategy to reduce computational complexity (Li et al., 2014; Louppe et al., 2016; Ventura et al., 2015).

A good blocking rule should maintain a balance between computational complexity and the maximum recall. Extent works that disambiguate inventor names written in alphabet usually block by surnames plus the first 1, first 3 characters of or the entire first names[18]. While it' is tempting to capture instances like typos or misspelling by blocking records with surnames in

[18] Studies of Milojević et al. (2013), Louppe et al. (2016) argues that blocking by the combination of surname plus first character of the first name could achieve higher disambiguating accuracy of papers. To patents, a narrower blocking strategy such as surname plus the first three characters of first name would be more ideal as inventors usually register their full names on patent documents. (Ventura et al., 2015)

Chinese characters plus the pinyin format of first names[19], the exponentially increased computational burden and the surging false positive errors (lumping errors) forced us to abandon this option. In line with Ikeuchi et al., (2017), we blocked records by inventors' exact full names and did not make any separation between first names and family names. This means all records with the same Chinese name could have a chance to be compared with each other within the same block while those with different names are deemed as different persons. Using this narrower blocking rule, the computational complexity decreases to 1.98 billion with a small rise in the false negative errors.

In SIPO inventor data, there are 490 common names referring to more than 1,000 patent records (7% of all records). Different from solutions that limit the maximum size of a block to be 1,000 (Ventura et al., 2015), we argue that no constraints like this should be imposed on Chinese names as this would incur vast splitting errors for those prolific inventors.

3.2.2 Building similarity Vector

With all primitive and derived information available, this step transforms inventor-patent records within a block into a similarity vector which compares their degree of similarity and then judges whether two inventor-patent records are match or non-match. This vector of similarity scores is called similarity profile, and it can be computed with heuristics like the number of items (e.g., coauthor names) shared or functions like Levenshtein distance, Jaro-Winker distance, cosine similarity, Euclidean distances, Manhattan distance, etc.

As mentioned above, Chinese patent dataset lacks the inventors' location information to distinguish inventors and track their mobility. Here we utilize the sole location information available: the address of the first applicant. In SIPO, raw data of applicant address is a string without being parsed into the hierarchical administrative level in China: i.e., the country –

[19] With such a blocking rule, intentional or unintentional name variants, typos like "但智刚" vs. "但智钢" could have a chance to be compared and be classified as the same person. However, names like "励士峰" and "励土峰" would have no opportunity to be compared thus splitting error arises.

province – city – district - road/village - road-number- room number. To calculate the of graphical similarity between two records, first, we fetched the latitude and longitude of applicants' locations in China from Baidu Map's public API, including their degree of confidence with their geocoding results. Then we input these latitude-longitude coordinates and received a 7-level parsed information detailed up to the exact building and room number level via their "reverse geocoding" services. Meanwhile, we also compare the Levenshtein distance of two records address in string format incase geocoding is wrong or when Baidu Map has low confidence on the result.

For inventors' affiliations, we compute the number of shared elements of two records' applicants as well as the string similarity of their first applicant (all by the stem of applicants names). To capture as much information about inventors' affiliation as possible, we also harmonized applicant's name with the string similarity of applicant name and their geographical location. For details of applicants' harmonization, please see the Appendix 2.

Although the IPC codes have covered information about the technological field that inventors relate to, it is a little general, and we assume that the concrete contents of an invention would also assist in the linkage decision. To the best of our knowledge, only Kim et al.'s (2016) work leveraged information about title. Whereas they use the number of words shared among two comparing records, we adopt the string similarity of two titles. Additionally, we also extracted two keywords from the title with the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm and compared their string similarity. For all the string comparisons we adopted the Levenshtein distance instead of Jaro-Winkler distance as the former could have better differentiating results for short text written in Chinese characters while the latter is more suitable for measuring distances of short texts in alphabet writing.

For the detailed definition of features, please refer to Table 6.

(Table 6)

3.2.3 Model Selection and Evaluation

This step combined similarity scores from many different dimensions into a single one and supervised learning approach usually sets the predicted probability for non-matchness as the distance between two records.

Before throw into machine learning classifiers, we performed some preprocessing like normalization of these features into [0,1] range to make sure the model uninfluenced by differences of the unit and range among features. In addition, as our training set is imbalanced (more 0 than 1), oversampling for training data would avoid the estimates being influenced by this uneven distribution and improve performances of some classifiers. To avoid overfitting, we also split our labeled data into training-validation-testing set and conducted all feature selection, model comparison and fine-tuning of hyper-parameters within training and validation set.

According to the no-free-lunch theorem in machine learning, there is no such model that always outperform all other models on all datasets. To find the best model of our disambiguation task, we tried seven kinds of supervised learning classifiers, i.e., Naïve Bayes models, logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) as well as 3 kinds of tree-based method, i.e., Random Forest, AdaBoost, and Gradient Boosting Decision Tree (GBDT). Other popular models like K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) were abandoned due to the long-training time and insignificant gain on performances.[20]

To test the reliability of the rare name data, we designed the following experiment with 5-fold & 10-fold cross-validation to evaluate the quality of training set and our model performance:

1) The rare name data as training set and 20% of hand-labeled data as testing set;
2) The 80% randomly selected data from hand-labeled dataset as training and the rest 20% as testing set;

[20] Our practice confirms opinions of previous studies in that classifiers like SVM (Ventura et al., 2015) and instance-based algorithms like KNN are inappropriate for disambiguation algorithms involving large-scale pairwise comparisons with high computational complexity.

3) The mix of rare name data and 80% randomly selected data in hand-training data as training set while the rest 20% as testing data.

This process is repeated 5 times until all points within hand-labeled data serve as the testing data only once. To avoid higher accuracy generated by records with the same name existing both within training and validating sets, we split record pairs according to their IDs (i.e., inventors' names).

We then applied the winner model trained on the selected training data (in our case, the combination of all rare name data with hand-labeled data) to the remaining unannotated data and predicate the possibility for two record pairs to be match and non-match. The probability for non-match was adopted as the distance between a pair and transformed into distance matrix for clustering records.

3.2.4 Clustering

Although much more popular, hierarchical linkage clustering performs poorly in our experiment. Due to the vast differences on shapes of dendrogram generated for records even with the same blocking size, we could not find an appropriate threshold that achieves a satisfying result for records with the same blocking size (let alone the global optimal) with hierarchical linkage clustering. Perhaps a threshold adaptive to the shape of clusters such as the "semi-supervised adaptive-height snipping of the hierarchical clustering tree" (Obulkasim, Meijer, & van de Wiel, 2015) would assist in solving this problem. On the contrary, clusters with arbitrary shapes are just where density-based clustering method like DBSCAN shines. For DBSCAN, it is unnecessary to pre-specify the number of clusters, easier to adjust parameters and it could form clusters according to the shape of clusters. Additionally, it allows the existence of noise, i.e., it would generate singular clusters with only one element that fits well with the fact that the majority of inventors only participated in the invention of one patent. Therefore, we choose DBSCAN to cluster records.

The last problem for clustering is choosing parameters for grouping the same individual into a cluster. There are two substantive parameters in DBSCAN: Eps (ε, the maximum radius of the neighborhood) and MinPts (the minimum number of points required to form a dense region). It is

these parameters that determine the final assignment of identifiers to records. Simply speaking, the larger the Eps, the large the lumping error and the smaller the splitting error, and vice versa. In searching for the best parameters, we tried both the supervised learning based on the rare name data and the unsupervised searching based on the Silhouette score, which is defined as

$$s = \frac{b - a}{\max(a, b)}$$

where a refers to the mean intra-cluster distance and b refers to the mean distance between a sample and all other points in the next nearest cluster. To find the optimal (range of) parameters, we first search within a very a broad range [0.01,0.99] with large steps and then gradually reduce the learning step.

3.3 Evaluation metrics

Evaluation of methods could be conducted at such four levels: inventors' names, individuals (predicated clusters of patent-inventor records) and pairs of patent-inventor records. To date, scholars have tried the individual-level or pairwise-level evaluation metrics like B-Cubed score (Louppe et al., 2016), error metrics like Type I (false positive) and II errors (false negatives) (Pezzoni et al., 2014), precision-recall rate (Raffo & Lhuillery, 2009) or their harmonic mean, the F1-score (adopted in PatentsView's contest). To achieve a low and balanced result between two types of errors, and to compare with previous studies, this paper adopted the more popular F1-score weighted by the number of instances for each label to choose and evaluate a model's performance, since the distribution of matching and non-matching cases is unbalanced. F1-score is defined as the harmonic mean of precision and recall rate:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} = \frac{TP}{TP + FP}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{TP}{TP + FN}$$

When weighted F1-score are too close to support the model selection decision, AUC score, which the area under the Receiver Operating Curve (ROC), would be also referenced since our testing data is unbalanced. At the same time, we also provide the splitting and lumping errors of our method. This intuitive is proposed by Tovik and Smalheiser (VETLE I. Torvik & Smalheiser, 2009) and then adopted by (Li et al., 2014) and thus Ikeuchi et al. (2017). According to Tovik and Smalheiser (VETLE I. Torvik & Smalheiser, 2009), splitting error refers to the error when an ID referring to a single person is split into many different inventor IDs and lumping error denotes the situation in which many different inventors are assigned with the same ID. Nonetheless, as an individual-level metrics, this error metric is defined according to the percentage of individuals erroneously mapped to the largest cluster of records. It only focuses on the largest cluster and thus fails to consider the number and size of all the clusters. Therefore, we adopt Ventura et al.'s (Ventura et al., 2015) revised version of Tovik and Smalheiser (VETLE I. Torvik & Smalheiser, 2009) and measure our errors at the pairwise level:

$$Splitting\ error = \frac{\#\ of\ pairs\ incorrectly\ as\ non-matches}{Total\ \#\ of\ true\ matches} = \frac{FN}{TP+FN}$$

$$Lumping\ error = \frac{\#\ of\ pairs\ incorrectly\ as\ matches}{Total\ \#\ of\ true\ matches} = \frac{FP}{TP+FN}$$

Apparently, the lumping error shows the prevalence of false positives, i.e., the situation in which we falsely classify records into positive (in our case, the match) and the splitting error represent that of false negative.

4. Results

4.1 Feature selection & Parameter tuning

After holding out 10% of names from hand-labeled data, we compared the weighted F1 score of the following seven models (keeping all hyper-parameters the same) with feature combination 1) 6 basic features: common member of applicants and co-inventors, similarity in IPC class and group, first applicant's address and string similarity of title; 2) 6 features in 1) plus two keywords

extracted from title; 3) 6 features plus two highly correlated features (i.e., app_i with app_s, geo with address_s); and 4) all 10 features.

(Figure 3)

The comparison results presented in Figure 3 indicate that, in general, models with more features could have better scores. Second, the effects of two keywords extracted from titles could improve the result, but their effect is much fewer compared to the string similarity of applicant and address. This is explained by the fact that most of the similarity score of keywords is 0. Third, for tree-based models, adding features correlated with existing ones would have complementary effects and thus improve the model, while this is not the case for other models. This is not surprising as unlike linear models and Naïve Bayes model which assume the independence of features, tree-based methods do not have such assumptions and they could handle correlated features efficiently.

(Figure 4)

The scaled Gini statistics for features' relative importance of GBDT in determining classification results is depicted in Figure 4. In conformity with our prediction, the number of shared applicants by two comparison pairs is the most critical feature, and the similarity of geolocation ranks the second. This is consistent with Ventura et al.'s (2015)findings. Keywords and IPC group have very limited contribution. In addition, the number of shared co-inventors did not as play as much importance as that in Monath & McCallum's work. This confirms our assumption that no feature has the deterministic influence on the classification result and when the shared member of co-inventors are two or more common names, it is possible that two records refer to two distinct persons.

4.2 Comparison of supervised models

Figure 3 already displays that GBDT yield the best result with different combinations of features. Result presented in Table 7 confirms this conclusion with a comparison of models with 5-fold cross-validation performed on three kinds of training-testing sets, while everything else (including the testing data) is kept the same. Based on these two tables, we summarize our model selection results as follows: first, among these models, Gradient Boosting classifier achieved the

highest scores across different training sets. In consequence, we would choose the GBDT classifier to predict the distance between two records in the final model due to its stable and extraordinary performance, insensitive to the scale and robustness to outliers.

(Table 7) & (Table 8)

To test the validity of this conclusion, we performed 10-fold cross-validation (Table 8) with different groups of names as testing data. Here the basic trend holds while the combination of two dataset yield slightly lower scores compared to purely using the rare name data as training set. What's more, there are also slight improvements as a result of increased training data and fewer testing data.

(Table 9)

Table 9 presents the detailed results of models when we use the rare name data as training set and 80% of the remaining hand-labeled data as the testing set. Results measured in AUC score is consistent with those measured with F1-score. Second, all models yield a much larger splitting error compared to lumping error (all above 10%). This is similar to the results of disambiguation work conducted on USPTO and EPO which means we are more prone to falsely split the same person into different IDs. As argued by Fegley & Torvik, (2013) and Pezzoni et al., (2014), large splitting error and smaller lumping error (i.e., high precision and low recall rate) would have a small negative impact on co-inventor network metrics or other classification of inventors compared to the opposite combination.

4.3 Evaluation of algorithms with different training data

The second significant point that Table 7 & 8 provide is: as training set, compared to the hand-labeled data with small size and biases, the more objective rare name data could be a qualified substitute for artificially annotated data. This justifies the reliability of the rare name data collected independently of patent information when the time-consuming hand-labeling could only result in small amount of data with biases. Furthermore, training models with a heterogeneous but larger dataset set (a combination of rare name data with 80% from hand-labeled dataset) could improve the result compared to only utilizing rare name data as the training set. This proves that, on the one hand, our model would benefit from more data, and on

the other hand, hand-collected data could complement the rare name data in providing information about those "difficult" cases close the non-linear decision boundary.

Scores reported above are only evaluation before the clustering step. To test the accuracy of our full algorithm, we split the hand-collected data into $Hand_{554}$ which composed of 554 fully-labeled names with $Hand_{remain}$. Here we retained the model (with all other hyper-parameters unchanged) with 3 kinds of training sets but clustered with different eps (with MinPts set as 1). As stated in Section 3.2.4, we tried both the supervised and unsupervised method to choose parameters of DBSCAN. The searching result based on rare name data suggest an eps of 0.11 between [0.10, 0.12] while Silhouette scores for some randomly selected samples suggest 0.0375 within the range of [0.03, 0.04]. The final decision should be determined by the authors according to their research topics. Here we present evaluation of our full algorithms with different eps in Table 10:

(Table 10)

First of all, this result on different training set seems to be in contradiction with results before clustering in that the $Hand_{remain}$ could have the best scores compared to rare name as training set. Here we argue that these results depend on methods of splitting and distribution of both training and testing data. Though a little lower, the rare name data still could yield results compatible with hand-labeled data as the training data.

Second, our searching of the optimal value of eps demonstrates we can achieve very high F1 scores (above 99%) at the expense of increasing lumping errors. While the F1-score rise above 99%, the AUC score declines below 90%. The main purpose of constructing this dataset is to serve inventor-level academic research, for which a large splitting error would not have huge influence on empirical results. Accordingly, though we take no preference for any type of errors in the classifying stage, we choose a very conservative threshold and prefer lowering lumping errors in final results to avoid generating too much "fake mobility".

Third, both F1-scores and AUC scores get higher while the splitting and lumping errors become much smaller compared to those evaluated before clustering. To conclude, based on all scores we received, we claim our algorithm could yield a satisfying F1-score ranging from

93.5%~99.3% (including before or after clustering) with splitting error within the range 0.5%~3.2% and lumping error 0.05%~0.37%.

5. Inventor mobility in China

From 1.8 million inventor names, we identified 3.99 million unique persons. Based on this result, here we present some preliminary findings on the regional mobility of Chinese inventors. In this section, we take all inventors with Chinese names (regardless of their location inside or outside of mainland China) whose location has changed at least once. Since there is no information about inventor address, we use the address information of their first applicant. In addition, it should be noted that the identification of mobile inventor depends on the granularity of location information. For example, an inventor moving accross cities may not be mobile ones at province level, since she moves within the same province.

Figure 5 shows the count of inventors by province, as well as their mobility across provinces. Beijing, Jiangsu, and Guangdong are top three provinces in terms of the numbers of inventors. It also exhibits that the numbers of inventors moving in and moving out are relatively larger in Beijing, as compared to Jiangsu and Guangdong. Shanghai, Shandong, and Zhejiang follow these top three provinces. Figure 6 displays the net number of inventors moving across provinces. Actually, large cities like Beijing and Shanghai are losing inventors (negative net numbers), while the provinces surrounding such cities, such as Jiangsu, Zhejiang, and Anhui, gain a lot. In addition, Sichuan and Fujian provinces show a relatively large gain, while northeast regions such as Liaoning, Jilin, and Heilongjiang are losing.

(Figure 5) and (Figure 6)

Moving forward, we focus on three cities, Beijing, Shanghai and Shenzhen, the three largest cities in terms of patent applications in China. These three cities are different in many aspects. Beijing is a capital city where many public research institutes and high-quality universities are located, while Shenzhen is a commercial city which has been developed as a production site of foreign firms based in Hong Kong (Mao & Motohashi, 2016). Shenzhen hosts major high-tech firms, such as Huawei, ZTE, BYD, and Tencent. Shanghai is also a large commercial city, where the presence of multinationals is relatively large (Dang et al., 2017).As revealed in Table 11, the

inventor mobility rate in Beijing and Shenzhen is more than 10%, while that of Shanghai is around 7%. In addition, the share of inventors working at start-up firms (defined by applicants of private firms that started to file patents after 2009) is relatively large in Shenzhen as compared to the other two cities.

(Table 11)

Table 12 looks at the matrix of inventor mobility including these three cities, the other places in China and outside China. Table 13 exhibits the share of the places of origin of inflowing inventors in three cities. Shenzhen attracts more people from Beijing and Shanghai, while the share of other places of origin for inventors in China is the largest in Beijing. In terms of foreign returnee, Shanghai has the largest share, followed by Shenzhen.

(Table 12) and (Table 13)

6. Discussion

6.1 Conclusion

Disambiguation of Chinese inventors has become the bottleneck of inventor-level studies and is widely accepted as a huge challenge for existing US-Centric or English-names-centric algorithms as Chinese names barely have the typical middle names or name variants, and its common name (homonym) problem is more severe than English names. This problem becomes increasingly urgent along with the fast expansion of Chinese authors and inventors recently. While some rule-based methods disambiguating Chinese authors of papers have been proposed (Han et al., 2017; Tang & Walsh, 2010), method utilizing machine learning approach based on their own characteristics is unavailable yet. In this paper, we created the first systematic framework for disambiguating names of applicants and inventors in SIPO database for academic and business analysis. Compared to the disambiguation of USPTO or EPO patent database, which emphasizes on linking different spellings into the same unique individual (the synonym problems) and usually assigns all people with the same name as the same person thereby create fake mobility, here we highlight that the “name game” played on Chinese inventors should focus on identifying different person with the same common name.

In this article, we choose the supervised learning approach because of its higher accuracy. For supervised learning, better training samples usually beat better algorithms and thus a high-quality training data is crucial. Learning from Li et al.(Li et al., 2014), Ventura et al. (2015) and Ikeuchi et al.(2017) , we overcome flaws in existing works and provide evidence for the reliability of labeled training data constructed from rare name list as a substitute for hand-labeled data when large-scale and representative labels are expensive, prone-to-error and even impossible to collect. Our results demonstrate that after several rounds of cleaning, an artificially generated data based on a list of rare names filtered from extra data sources could provide even better classification results compared to the hand-labeled data which might be more accurate but much smaller in size and containing larger bias. A combination of the real and generated datasets in which both have defects could have complementing effects and improve the result. This evidence is not only beneficial for disambiguating East-Asian names but could also be applied to disambiguating names of other countries where name information in national level census data or basic information of citizens are public or easier to access. While being able to produce accurate, large-scaled and low-biased training set quickly, this method is also easy to update[21] and covers complex situations of inventors like regional and international mobility, participating in inter-disciplinary inventions or completely changing the industry they work within.

Scores received from 5 and 10-fold cross-validation manifest that Gradient Boosting classifier have the best and most stable performances across different training sets with different combination of features. Therefore, it was chosen to train the model and predict distances between records. We also propose that DBSCAN outperform hierarchical clustering in clustering records on the SIPO patent dataset in its simplicity, scalability to blocks with varying size and flexible shapes.

Evaluation before and after clustering demonstrates our algorithm can yield a satisfying F1-score ranging from 93.5%~99.3% with final splitting error within the range 0.5%~3.2% and

[21] Update the manually-annotated data means relabeling each inventor again thus pose another severe problem for training data within supervised learning, as pointed out by Ferreira, Gonçalves, & Laender, (2012)

lumping error between 0.05%~0.37%, depending on the testing set and different parameters of clustering. We believe this result is good enough be adopted in academic research.

6.2 Limitations and future directions

We only extracted and disambiguated Chinese inventors and left Non-Chinese names unprocessed. But disambiguation in practice are usually mixed of inventors or scholars from various ethnic or cultural backgrounds. For such kind of tasks, training a classifier to predict the ethnicity of inventors and adding these extra features of ethnicity may provide a better solution. Second, as mentioned previously, though we have tried to control the biases in the hand-labeled dataset and make it as random and representative as possible, it still contains biases like self-reporting bias, bias toward prolific inventors and university professors. Third, considering our blocking strategy rules out situations like name changing, typos, etc., our actual error rate would be slightly larger than reported in this paper. Fourth, without extra information, it is hard for human beings and computers to distinguish records with the same name and affiliation given the data available because neither their location nor technological dimension (classifications or contents) could have large differences. Therefore, for inventors from large companies like China National Petroleum Corporation (中国石油天然气集团) which has 1,512,048 employees, Huawei about 180,000 personnel[22] or even large and famous universities like Zhejiang University, the lumping error would be larger than those from smaller applicants. Features representing the size of applicant measured by the number of employees or number of patents filed might contribute to solving this problem (Torvik & Smalheiser, 2009). Finally, as disambiguation work usually consists of a small portion of labeled data and a vast majority of unlabeled ones, semi-supervised learning, which makes use of the distribution of unlabeled data, also has great potential. It's worthwhile to be tried in the future.

[22] All obtained from http://www.fortunechina.com/fortune500/c/2017-07/20/content_286807.html and Huawei's official website in 2017.

Acknowledgement

This work is mainly supported by The Research Institute of Economy, Trade and Industry's (RIETI) under the project of Empirical Analysis of Innovation Ecosystems in Advancement of the Internet of Things (IoT), NSFC-JSPS Scientific Cooperation Program between China and Japan (No.71711540044) and National Natural Science Foundation of China (No. 71503123). We also appreciate Dr. Ikeuchi Kenta's insightful suggestions.

References

Akinsanmi, E. O., Fuchs, E., & Reagans, R. E. (2011). Economic Downturns, Technology Trajectories and the Careers of Scientists. Retrieved from https://smartech.gatech.edu/handle/1853/42529

Azoulay, P., Zivin, J. S. G., & Sampat, B. N. (2011). *The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine* (Working Paper No. 16683). National Bureau of Economic Research. https://doi.org/10.3386/w16683

Balsmeier, B., Chavosh, A., Li, G. C., Fierro, G., Johnson, K., Kaulagi, A., ... & Fleming, L. (2015). Automated disambiguation of us patent grants and applications. *Fung Institute for Engineering Leadership Unpublished Working Paper*.

Cassi, L., & Carayol, N. (2009). Who's Who in Patents. A Bayesian approach. Retrieved from https://hal-paris1.archives-ouvertes.fr/hal-00631750/document

Chin, W.-S., Zhuang, Y., Juan, Y.-C., Wu, F., Tung, H.-Y., Yu, T., … Lin, C.-J. (2014). Effective String Processing and Matching for Author Disambiguation. *Journal of Machine Learning Research*, *15*, 3037–3064.

Dang, J., Mao, H., and Motohashi, K. (2017). Physically Proximate or Culturally Cohesive? Geography, Ethnic Ties and Innovation in China. *Working Paper*.

Fegley, B. D., & Torvik, V. I. (2013). Has Large-Scale Named-Entity Network Analysis Been Resting on a Flawed Assumption? *PLOS ONE*, *8*(7), e70299. https://doi.org/10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Rec.*, *41*(2), 15–26. https://doi.org/10.1145/2350036.2350040

Fleming, L., King, C., & Juda, A. I. (2007). Small Worlds and Regional Innovation. *Organization Science*, *18*(6), 938–954. https://doi.org/10.1287/orsc.1070.0289

Giles, C. L., Zha, H., & Han, H. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)* (pp. 334–343). https://doi.org/10.1145/1065385.1065462

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools* (Working Paper No. 8498). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w8498

Hall, B. H., Thoma, G., & Torrisi, S. (2007). The Market Value of Patents and R&d: Evidence from European Firms. *Academy of Management Proceedings*, *2007*(1), 1–6. https://doi.org/10.5465/AMBPP.2007.26530853

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries.* (pp. 296–305). https://doi.org/10.1109/JCDL.2004.240051

Han, H., Xu, W., Zha, H., & Giles, C. L. (2005). A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (pp. 1065–1069). New York, NY, USA: ACM. https://doi.org/10.1145/1066677.1066920

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics*, *111*(3), 1879–1896. https://doi.org/10.1007/s11192-017-2338-6

He, Z., Tong, T., Zhang, Y., & He, W. (2017a). Construction of a database linking SIPO patents to firms in China's Annual Survey of Industrial Enterprises 1998-2009. *Working Paper*. Retrieved from https://sites.google.com/site/sipopdb/home/sipo---asie

He, Z., Tong, T., Zhang, Y., & He, W. (2017b). SIPO - Chinese listed firms - Chinese Patent Data Project. *Journal of Economics & Management Strategy*. Retrieved from http://dx.doi.org/10.7910/DVN/CF1IXO

Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient Name Disambiguation for Large-Scale Databases. In *Knowledge Discovery in Databases: PKDD 2006* (pp. 536–544). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11871637_53

Ikeuchi, K., Motohashi, K., Tamura, R., & Tsukada, N. (2017). Measuring Science Intensity of Industry using Linked Dataset of Science, Technology and Industry Discussion. Dissucssion papers 17056, Research Institute of Economy, Trade and Industry (RIETI).

Khabsa, M., Treeratpituk, P., & Giles, C. L. (2014). Large scale author name disambiguation in digital libraries. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 41–42). https://doi.org/10.1109/BigData.2014.7004487

Kim, K., Khabsa, M., & Giles, C. L. (2016). Inventor name disambiguation for a patent database using a random forest and DBSCAN. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (pp. 269–270).

Lai, R., D'Amour, A., & Fleming, L. (2011). *The careers and co-authorship networks of U.S. patent-holders, since 1975*. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/12367

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., … Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy*, *43*(6), 941–955. https://doi.org/10.1016/j.respol.2014.01.012

Lissoni, F., Sanditov, B., & Tarasconi, G. (2006). *The Keins Database on Academic Inventors: Methodology and Contents* (KITeS Working Paper No. 181). KITeS, Centre for Knowledge, Internationalization and Technology Studies, Universita' Bocconi, Milano, Italy. Retrieved from https://econpapers.repec.org/paper/cricespri/wp181.htm

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). Ethnicity Sensitive Author Disambiguation Using Semi-supervised Learning (pp. 272–287). Presented at the International Conference on Knowledge Engineering and the Semantic Web, Springer, Cham. https://doi.org/10.1007/978-3-319-45880-9_21

Mao, H., & Motohashi, K. (2016). A Comparative Study on Tenant Firms in Beijing Tsinghua University Science Park and Shenzhen Research Institute of Tsinghua University. *Asian Journal of Innovation & Policy*, *5*(3), 225–250. https://doi.org/10.7545/ajip.2016.5.3.225

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, *7*(4), 767–773. https://doi.org/10.1016/j.joi.2013.06.006

Morrison, G., Riccaboni, M., & Pammolli, F. (2017). Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific Data*, *4*. https://doi.org/10.1038/sdata.2017.64

Müller, M.-C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: an empirical analysis and a new resource. *Scientometrics*, *111*(3), 1467–1500. https://doi.org/10.1007/s11192-017-2363-5

Obulkasim, A., Meijer, G. A., & van de Wiel, M. A. (2015). Semi-supervised adaptive-height snipping of the hierarchical clustering tree. *BMC Bioinformatics*, *16*, 15. https://doi.org/10.1186/s12859-014-0448-1

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H. F., Gonçalves, M. A., & Ferreira, A. A. (2009). Using Web Information for Author Name Disambiguation. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 49–58). New York, NY, USA: ACM. https://doi.org/10.1145/1555400.1555409

Pezzoni, M., Lissoni, F., & Tarasconi, G. (2014). How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation. *Scientometrics*, *101*, 477–504. https://doi.org/10.1007/s11192-014-1375-7

Raffo, J., & Lhuillery, S. (2009). How to play the "Names Game": Patent retrieval comparing different heuristics. *Research Policy*, *38*(10), 1617–1627. https://doi.org/10.1016/j.respol.2009.08.001

Singh, J. (2005). Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science*, *51*(5), 756–770. https://doi.org/10.1287/mnsc.1040.0349

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, *84*(3), 763–784. https://doi.org/10.1007/s11192-010-0196-6

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805000/

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, *56*(2), 140–158. https://doi.org/10.1002/asi.20105

Trajtenberg, M., Shiff, G., & Melamed, R. (2006). *The "Names Game": Harnessing Inventors' Patent Data for Economic Research* (Working Paper No. 12479). National Bureau of Economic Research. https://doi.org/10.3386/w12479

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications Using Random Forests. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 39–48). New York, NY, USA: ACM. https://doi.org/10.1145/1555400.1555408

Ventura, S. L., Nugent, R., & Fuchs, E. R. H. (2015). Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, *44*(9), 1672–1701. https://doi.org/10.1016/j.respol.2014.12.010

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, *93*(2), 391–411. https://doi.org/10.1007/s11192-012-0681-1

Zhang, G., Guan, J., & Liu, X. (2014). The impact of small world on patent productivity in China. *Scientometrics*, *98*(2), 945–960. https://doi.org/10.1007/s11192-013-1142-1

Appendix 1 Results of existing disambiguation work on inventors of patents

(Table 14)

Appendix 2 Applicant standardization

Applicant name standardization is an obstacle that nearly all researchers would encounter if they desire to make use of Chinese patent data or link it with external data sources. Since this is also a disambiguation or record linkage task, thus our framework on disambiguating inventors could easily transplant to the harmonization of applicants' names. However, instead of suffering from the common name problem, what the harmonization of applicant name has to deal with is the synonym problem: its goal is to cover the name variant, changing of firm names and typos of the same applicant as accurate as possible.

We harmonized Chinese applicant names according to their characteristics: i.e., a typical Chinese Firm name is usually composed 4 parts—Province/city + name stem+ industry+ type (e.g., 深圳/TCL/数字技术/有限公司). While purely relying on the name or geocoding could go wrong, high similarity in both 2 dimensions indicates higher probability of matching. Here we compared both the rule-based and supervised learning method to standardize firm and university names. Although supervised learning method performs a little better than rule-based, the rule-based approach is applicable due to its simplicity and empirical studies usually do not have such high requirement for accuracy. The steps are as follows:

1) **Preprocessing:** removing spaces and name suffix like "股份有限公司", "有限公司", "(北京)", "研究所", "研究院" while keeping address prefix "北京", "深圳", "大学". This is because for many applicants, especially universities (e.g. "北京大学", "浙江大学", etc.), the address prefix is their sole identifier.
2) **Stemming:** extract the most special words in a name with TF-IDF algorithm: "万达", "中兴通讯", "TCL", "ABB". For wrong keywords which extracted industry mistakenly, some manual check and extraction by places and length of names are involved.

3) **Block by the stem, applicant type and generating comparing pairs:** to make sure comparison happens only among applicants with the same word stem and type. From 0.380 million unique applicant names, we generated 5.54 m record pairs to compare.
4) **Comparing record pairs:** to link records in or between data sources.
5) **Cluster with thresholds directly or predicating distances based on trained models and clustering with distance matrices (similar to steps in inventor disambiguation).** We collected three datasets of SIPO's patent data linked with external firm information for training and testing the algorithms: one is linked to firms on National Equities Exchanges and Quotations (NEEQ) with our manually disambiguated results, the second is SIPO's linked data with Chinese listed firms ("Main Board") and the last is linked with the Annual Survey of Industrial Enterprises (ASIE) (He et al., 2017a, 2017b). The latter two datasets is provided the Chinese Patent Data Project (CPDP) [23].

Appendix 3 Comparison of our definition of similarity profile with existing work

Since our task has more common points with the disambiguation of Japanese names, here we present differences between our definition of features with that of Ikeuchi et al. (2017):

(Table 15)

[23] These two datasets could be downloaded from https://sites.google.com/site/sipopdb/home/SIPO_listed and https://sites.google.com/site/sipopdb/home/sipo---asie

Tables & Graphs

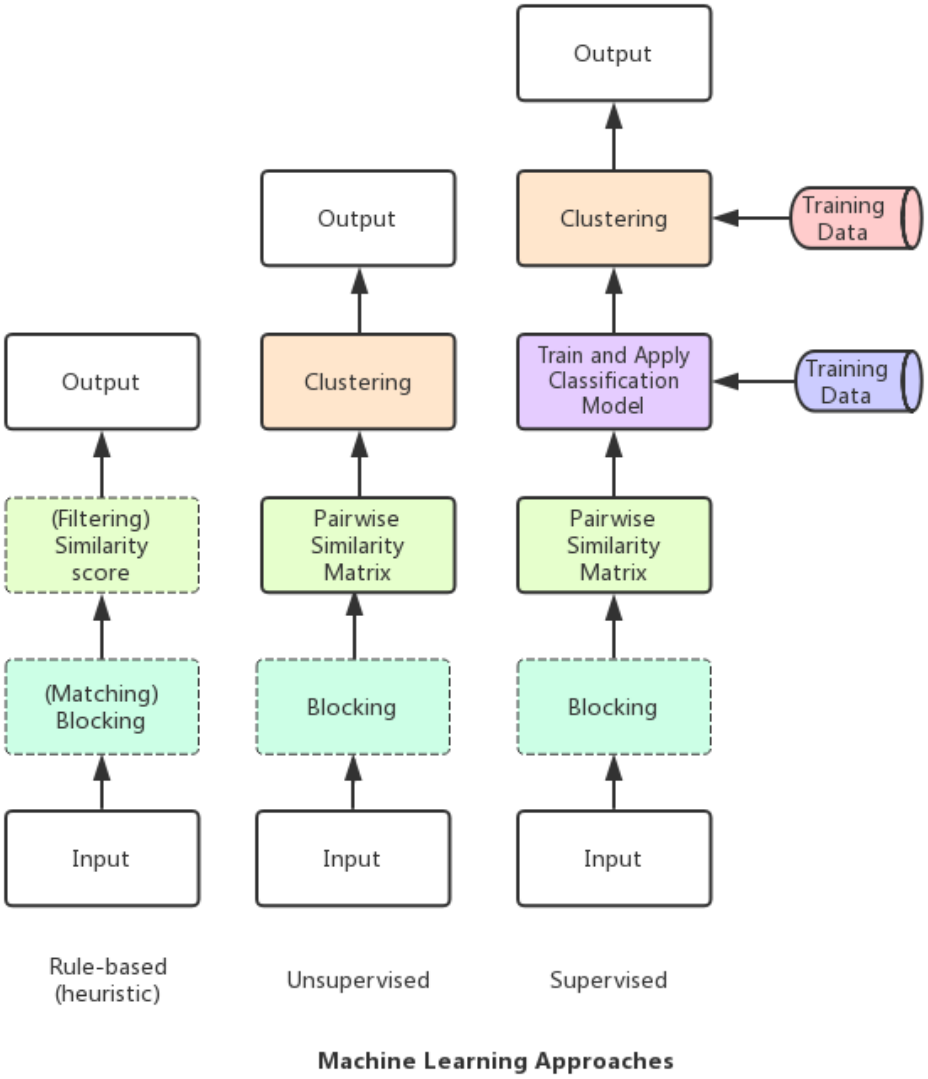
Table 1: Statistics and distribution of the blocking size in SIPO, JPO and USPTO

Blocking size[24]	SIPO To 2016.12		JPO To 2014.3		USPTO (last+ first 3) To 2017.08	
1-100	10,458,586	71.27%	10,309,875	83.16%	11,572,770	77.36%
100-500	2,637,172	17.97%	1,970,057	15.89%	2,413,536	16.13%
500-1000	616,617	4.20%	83,038	0.67%	469,805	3.14%
1000-max	963,109	6.56%	34,850	0.28%	503,541	3.37%
Total	14,675,484		12,397,820		14,959,652	

Source: Authors' calculations based on SIPO, IIP and PatentsView datasets.

Figure 1 Our taxonomy and classification of methodologies

[24] Here the blocking size refers to the number of patent one name corresponding to. Here names within SIPO and JPO are blocked by full names while those in USPTO is blocked by the combination of last name with first 3 characters of first name.



Source: Authors

Table 2: Literature review: existing works on the disambiguation of patent inventors

Methodology	Author	Application	Training Set	Algorithms	Evaluation
Rule-based (Heuristic)	Hall, 2001	All applicants in USPTO	NA	String match of assignees	First public dataset of disambiguated assignees
	(Zhang et al., 2014)	SIPO 2000-2009	NA	String matching of applicant, province, IPC class	First and unique attempt on disambiguating Chinese inventors available
	Pezzoni et al., 2014	PATSTAT 2011	NA	Similarity profile + ad hoc threshold and weight	First systematic disambiguation of PATSTAT data
	Morris, 2017	PATSTAT 2014	NA	using high-resolution geocoding data	Simple, straightforward rules that disambiguate assignee and inventors at the same time
Semi-supervised	Li & Lee Fleming, 2014	Full USPTO	Statistically generated labels with rare names as a part of input data	Naïve Bayes + ad hoc rules automatically generated training sets	Pioneering work based on machine-learning method
	Ikeuchi et al., 2017	Japanese inventors in JPO	Rare name data as a part of input data	Naïve Bayes + ad hoc rules	First systematic disambiguation of Japanese patent data
Supervised	Ventura, et.al., 2015	Small subset of USPTO	Optoelectronics (OE) + Academic life scientists (ALS)	Random Forest + Hierarchical clustering	First supervised method while the training set belongs to one industry and have large bias when applying to whole USPTO dataset
Unsupervised	Balsmeier et al., 2016	Weekly updated Full USPTO	NA	K-Means clustering	Completely automated process

Source: Authors

Table 3: Summary of nationality identified by names in full SIPO patent inventor data

Names of country	Unique patents	Unique names	Inventor-patent pairs	Ratio %	Ida_seq /names	Max	Criteria
Chinese (Korean, Taiwanese, etc.)	4.9 m	1.84 m	14,685,617	78.90	7.96	9680	No points within name, do not have a Japanese Family name and the length of name equal to or lower than 4 characters
Western (all other countries)	0.90 m	1.17 m	2,492,036	13.39	2.14	509	With a point in names, e,g., "P·T·贾特", "D·罗布"
Japanese	0.57 m	0.35 m	1,435,067	7.71	4.05	1402	With typical Japanese Family name e.g. "伊藤彰浩", "毒岛真"
In Total	6.25 m	3.3 m	18,612,720	100			

Source: Authors' calculation based on SIPO

Table 4: Representativeness measured by the type of records and statistics of hand-labeled data and rare name data

	Mean of patent-owned	Hand-labeled labeled	Hand-labeled & unlabeled under the same name	Rare names	All records remained
Individual	57.19	2.03%	6.21%	10.82%	9.44%
Firm	32.68	23.68%	53.31%	54.22%	57.55%
University	92.01	60.48%	29.54%	24.73%	22.99%
Research Institute	62.27	13.80%	10.94%	10.23%	10.02%
Total number		21,073	128,753	402,339	14,144,365
Median		56	16	3	4
Mean		86.45	138.59	6.073	12.66
Skewness		2.225	8.13	13.77	49.96
Kurtosis		8.655	84.56	377.86	4492.06
Range		[2,484]	[2,7167]	[2,635]	[2,9680]
Std. Dev.		94.43	528.52	12.17	57.61

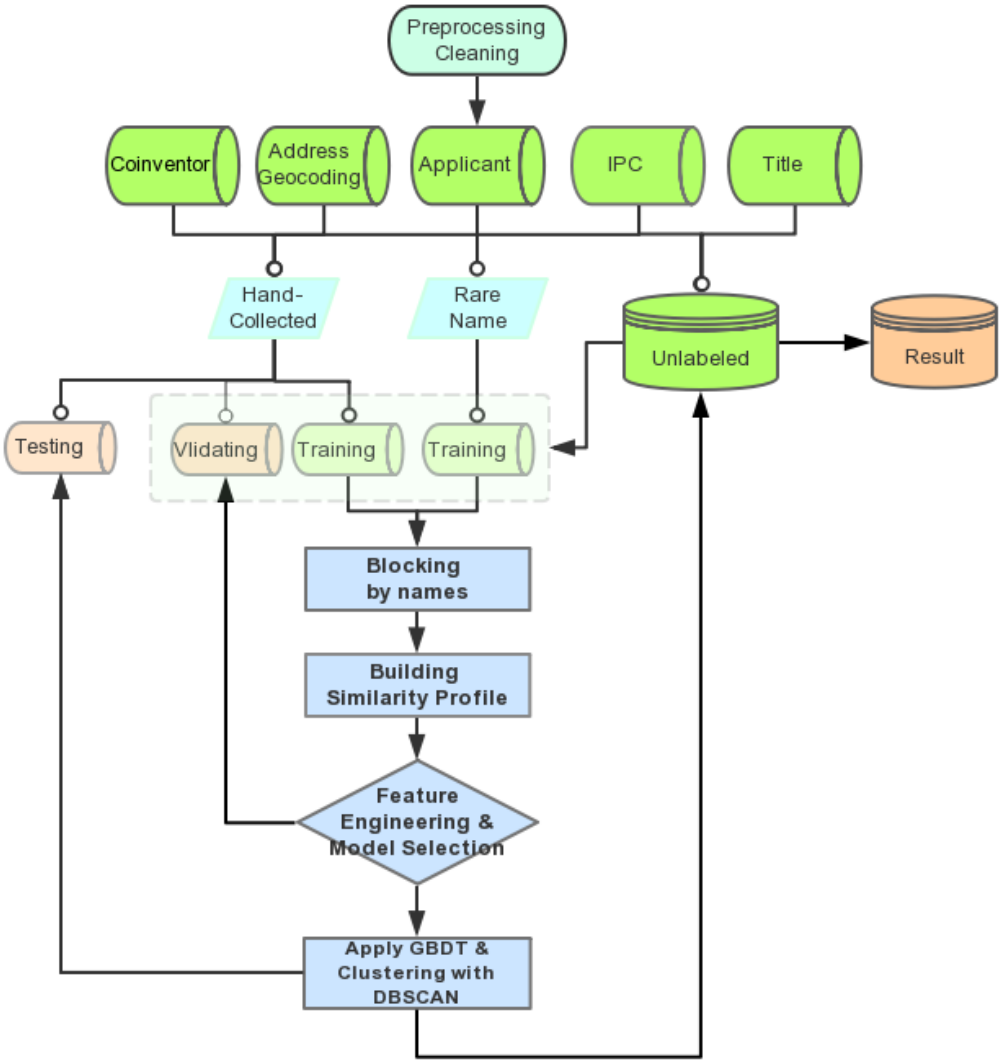
Source: Authors' calculation based on SIPO

Table 5: Statistics of hand-collected data, rare name data and all Chinese inventors in SIPO

	Hand-labeled data	Rare name data	SIPO Chinese inventors
# of unique person	1,314	66,248	NA
# of unique names	929	66,248	1.84 million
# of unique patents	16181	376,429	4.93 million
# of inventor-patent records	21,073	402,339	14.68 million 13.97 million
# of inventor-patent records with unlabeled	128,753	402,339	14.68 million 13.97 million
# of comparison pairs	1.73 million	12.27 million	1.98 billion
# of non-matches 0	0 1,099,550	0 6,133,152	NA
# of matches 1	1 629,993	1 5,927,216	

Source: Authors' calculation based on SIPO

Figure 2 Flowchart of our algorithm



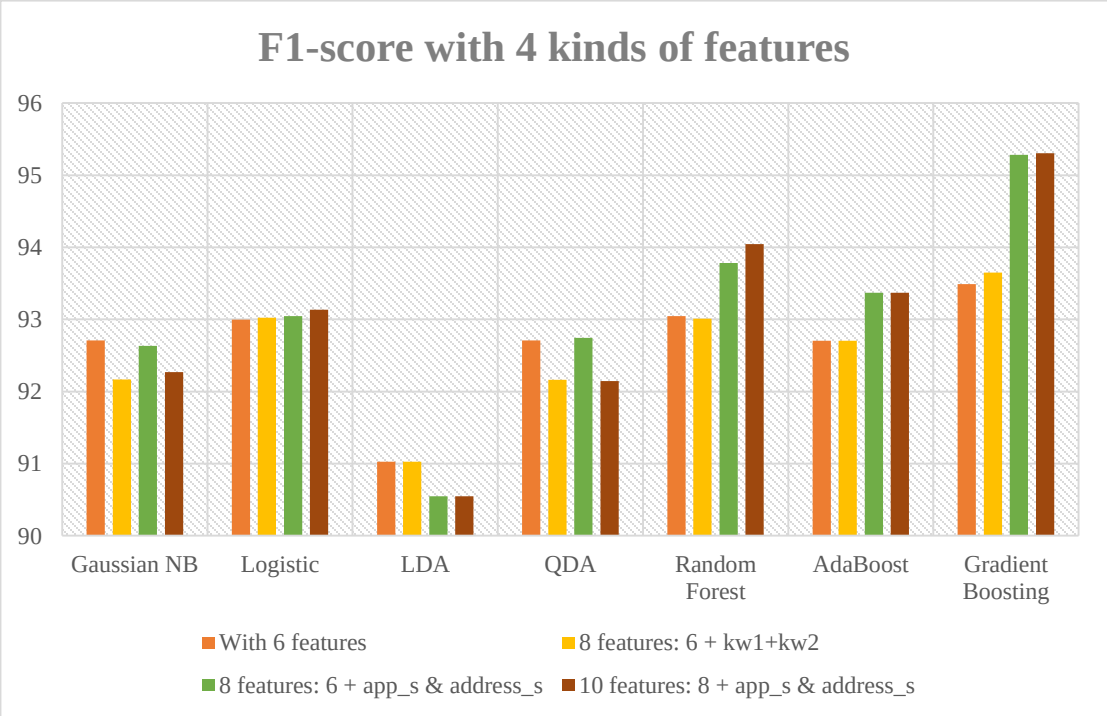
Source: Authors

Table 6: Definition of features in similarity profile

Feature groups	Feature	Definition
Applicants	app_i	# of common members of applicants
	app_s	String similarity of first applicants' names
Co-inventors	inventor_i	# of shared inventor's names of two records
Technological fields	ipc_c	# of common members of IPC class
	ipc_g	# of common members of IPC group
Technological content of patents	title	String similarity of titles
	keyword 1	String similarity of first keyword
	keyword 2	String similarity of second keyword
Address	address_s	String similarity of addresses
	geo	0 from a different country
		1 from the same country
		2 from the same province
		3 from the same city
		4 from the same district
		5 from the same road or village
		6 have the same latitude and longitude

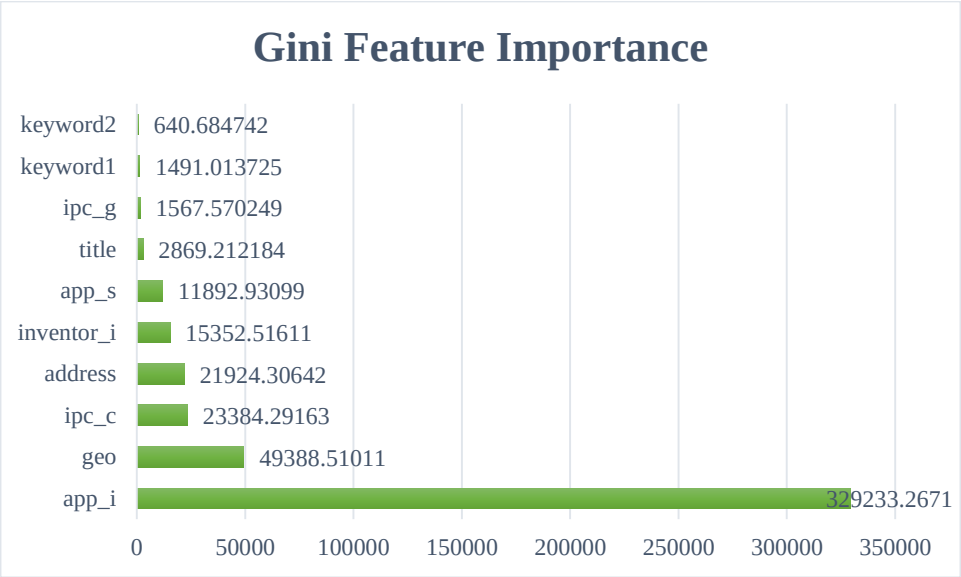
Source: Authors

Figure 3: Feature selection: F1-score received with 4 kinds of combination of features



Source: Authors' calculations based on our algorithm.

Figure 4: Gini importance of features in Gradient Boosting Classifier



Source: Authors' calculations based on our algorithm.

Table 7: Pairwise F1-score from 5-Fold Cross-validation and comparison of models with 3 kinds of training sets (Evaluation before clustering)

Training set / Models	(1) 80% of Hand-collected	(2) Rare name	(3) Mixed training set: Rare name + 80% of hand-collected
Naïve Bayes	92.23557	92.085072	92.173541
Logistic Regression	92.54353	92.808226	89.980191
LDA	90.44477	90.122986	92.84446
QDA	91.94083	92.052651	92.054074
Random Forest	91.69556	91.817041	92.324427
AdaBoost	92.94519	91.561233	91.799894
Gradient Boosting	93.13849	93.285912	93.362207

Source: Authors' calculations based on our algorithm.

Table 8: Pairwise F1-score from 10-Fold Cross-validation and comparison of models with 3 kinds of training sets (Evaluation before clustering)

Training set / Models	(1) 80% of Hand-collected	(2) Rare name	(3) Mixed training set: Rare name + 80% of hand-collected
Naïve Bayes	89.64731	92.34737	92.44074
Logistic Regression	90.76750	93.10603	93.10967
LDA	88.59201	90.41615	90.27762
QDA	89.40703	92.31727	92.32215
Random Forest	89.46714	92.22502	92.69485
AdaBoost	89.24354	91.81432	92.35146
Gradient Boosting	91.25289	93.48601	93.44492

Source: Authors' calculations based on our algorithm.

Table 9: Pairwise Score with rare name + 80% of hand-collected as training set (Evaluation before clustering)

	F1 Score	AUC Score	Splitting Error	Lumping Error
Naïve Bayes	92.173541	90.97826	14.88665	2.978167
Logistic	89.980191	87.2324	25.21166	0.286963
LDA	92.84446	91.26537	16.4657	0.944021
QDA	92.054074	90.96794	14.47707	3.408231
Random Forest	92.324427	91.22295	14.03578	3.318383
AdaBoost	91.799893	90.31288	17.05161	2.183577
Gradient Boosting	93.362207	92.36092	12.93609	2.254097

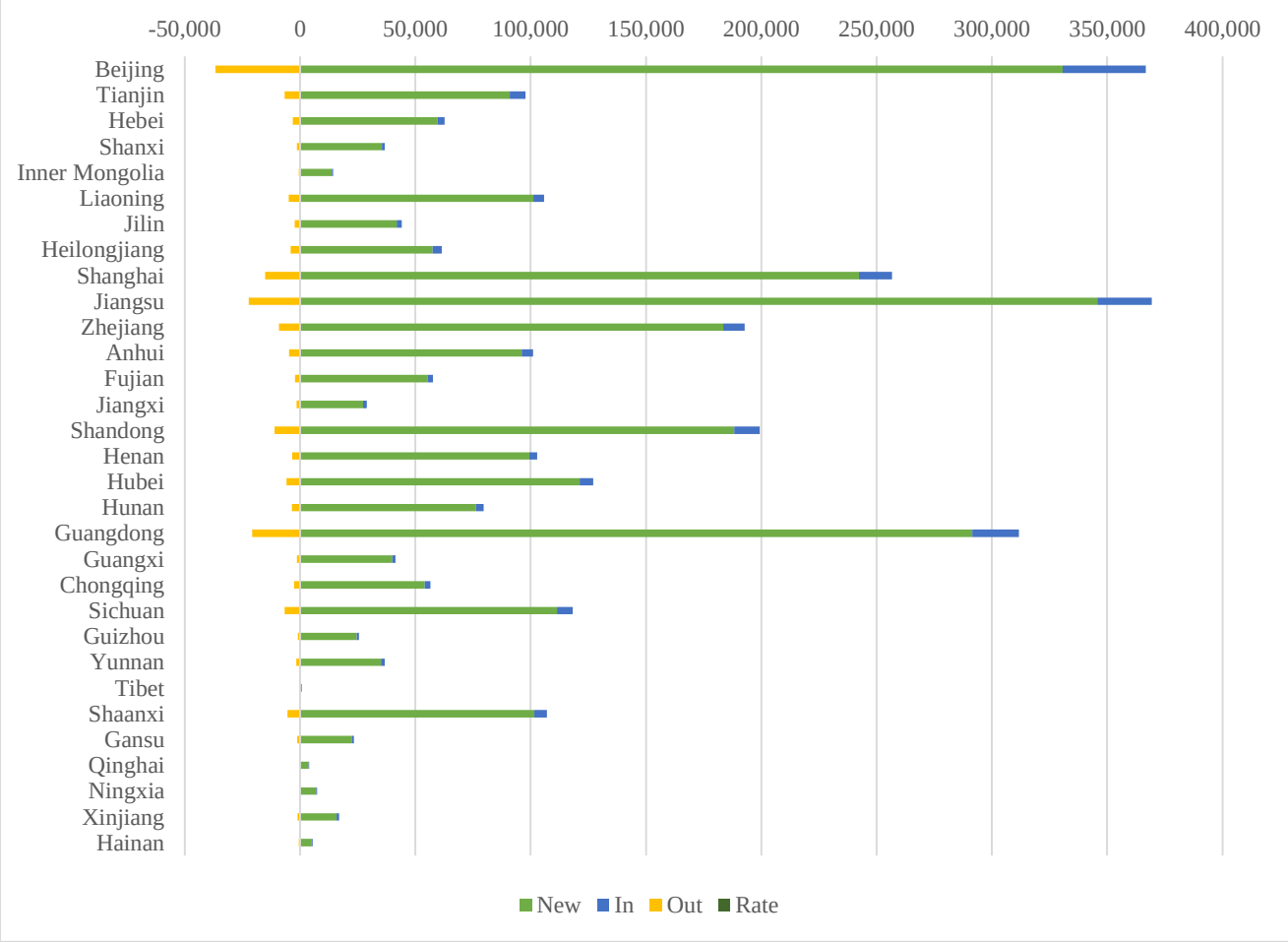
Source: Authors' calculations based on our algorithm.

Table 10: Comparison of models with 3 kinds of training sets (Evaluation after clustering)

EPS	Training set	Testing set	F1 Score %	AUC Score %	Splitting Error %	Lumping Error %
0.0375	**Rare**	$\text{Hand}_{\text{full554}}$	97.5844	96.4675	3.1765	0.05071
0.11	**Rare**	$\text{Hand}_{\text{full554}}$	97.6336	96.7583	3.0938	0.05405
0.0375	$\textbf{Hand}_{\textbf{remain}}$	$\text{Hand}_{\text{full554}}$	97.7592	95.9661	2.8564	0.08310
0.11	$\textbf{Hand}_{\textbf{remain}}$	$\text{Hand}_{\text{full554}}$	97.7129	95.8715	2.9262	0.08500
0.375	$\textbf{Hand}_{\textbf{remain}}$	$\text{Hand}_{\text{full554}}$	99.1963	89.8351	0.6493	0.31381
0.55	$\textbf{Hand}_{\textbf{remain}}$	$\text{Hand}_{\text{full554}}$	99.1133	88.2222	0.5602	0.36669
0.0375	**Rare + Hand**$_{\textbf{remain}}$	$\text{Hand}_{\text{full554}}$	97.6642	95.8770	3.0048	0.08357

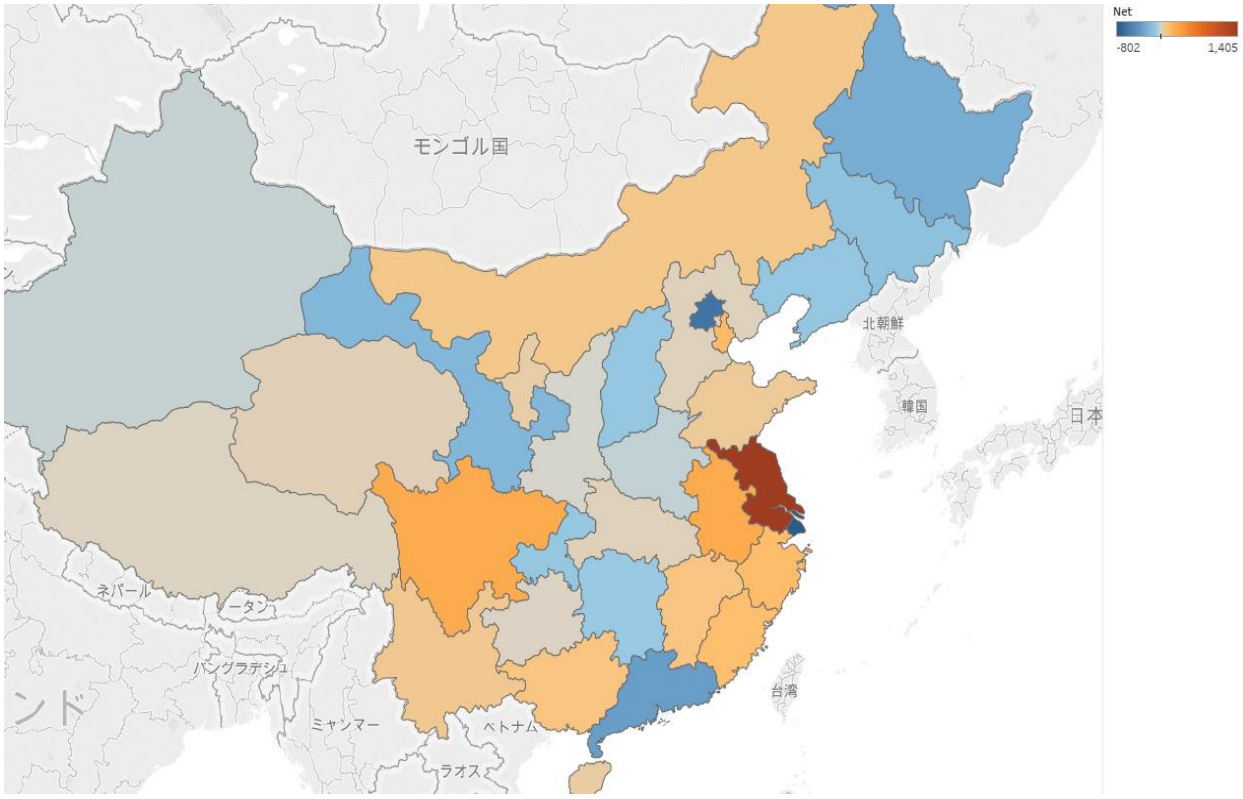
Source: Authors' calculations based on our algorithm.

Figure 5: Number of inventors at province level



Source: Authors' calculations based on our disambiguation result.

Figure 6: Net flows of inventors moving across provinces



Source: Authors' calculations based on our disambiguation result.

Table 11: Moving ins and outs for three cities

	Initial	From outside	To outside	Current	Share of Startup
Beijing	330,681	36,323	36,951	330,053	21%
Shanghai	242,091	14,363	15,100	241,354	24%
Shenzhen	128,197	15,114	15,438	127,873	30%

Source: Authors' calculations based on our disambiguation result.

Table 12: Mobility matrix of three cities

From/To	Beijing	Shanghai	Shenzhen	Other China	Foreign	Total	% of move out	# of inventors
Beijing	0	2,802	4,229	29,910	10	36,951	10.8%	341,606
Shanghai	2,832	0	1,398	10,861	9	15,100	6.1%	247,644
Shenzhen	4,209	1,385	0	9,803	41	15,438	11.6%	132,976
Other China	29,264	10,166	9,344	0	111	48,885	2.2%	2,198,947
Foreign	18	10	69	155	0	252	0.1%	276,503
Total	36,323	14,363	15,040	50,729	171			
% of move-in	10.6%	5.8%	11.3%	2.3%	0.1%			
# of inventors	341,606	247,644	132,976	2,198,947	276,503			

Source: Authors' calculations based on our disambiguation result.

Table 13: The place of origin of moving in inventors for start-up firms

From/To	Beijing	Shanghai	Shenzhen
Beijing	-	19.5%	28.1%
Shanghai	7.8%	-	9.3%
Shenzhen	11.6%	9.6%	-
Other China	80.6%	70.8%	62.1%
Foreign	0.0%	0.1%	0.5%

Source: Authors' calculations based on our disambiguation result.

Table 14: Results of existing disambiguation work on inventors

Authors	Dataset	Methodology and models	claimed Splitting Errors25	claimed Lumping Errors	F1 Score
Fleming et al., 2007 Lai et al., 2009	USPTO	Rule-based	Precision: 96.1%	Recall: 97.3%	
Morris et al., 2017	EPO, PCT, USPTO	Rule-based based on high-resolution geocoding data	10.5%	9.5%	
Pezzoni et al., 2014	PATSTAT	Rule-based	Precision: 88%	Recall: 68%	
Ventura et.al., 2015	Subset of USPTO	Optoelectronics + Academic life scientists; Random Forest + Hierarchical clustering	2.09%	1.26%	
Li et al., 2014	USPTO	Naïve Bayes + ad hoc rules with automatically generated training sets	3.26%	2.34%	3rd party's result: 92.7314%
Kim et al., 2016		Random Forest + DBSCAN	Precision: >99%	Recall: 97%	98.37%
Monath &McCallum, 2017		Graph-based method			98.16%
Ikeuchi et al., 2017	JPO	Naïve Bayes + ad hoc rules	2.41%	0.29%	
Balsmeier et al., 2016	USPTO	K-Means clustering			

[25]In this table we document authors' self-reported scores or error rates just for references. Nonetheless, no judgements about quality of these methods should be made based on this table as they work on different dataset and evaluated with different testing data of different size, as well as different evaluating methods.

Table 15: Comparison of our definition of similarity profile with that of Ikeuchi et al. (2017)

	Ikeuchi, et al., 2017	Ours
Inventor name	1 if names are completely same. 0 otherwise.	The same
Applicants	3 if applicant identification numbers are equal. 2 if applicant names are same. 1 if either applicant identification number or applicant name are not available. 0 if both applicant identification numbers and names are different.	# of shared members of harmonized applicant names # of shared members of applicant names String similarity of first applicants' names
Address	5 if matched at land number extension (go-level). 4 if matched at land number (banchi-level). 3 if matched at city block (chimei-level). 2 if matched at municipality-level. 1 if matched at prefecture-level. 0 otherwise.	string (address) 6 Latitude-longitude 5 Sub road/village level 4 road/village 3 district 2 city-level 1 province-level 0 Country
Co-inventors' names	# of shared co-inventors, where more than 6 common co-inventors is set to a maximum value of 6.	# of shared inventors, but no Maximum
Technology class	4 if main IPCs are same at 4 digit level. 3 if main IPCs are same at 3 digit level. 2 if main IPCs are same at 1 digit level. 1 if main IPCs are not available. 0 if main IPCs are completely different.	# of shared IPC class # of shared IPC group
Title	NA	String similarity of whole title String (keyword 1) String (keyword 2)

Source: Authors and Ikeuchi et al., 2017