Construction of a Database Linking SIPO Patents to Firms in China's Annual Survey of Industrial Enterprises 1998-2009

Zi-Lin He
Tilburg School of Economics and Management
Tilburg University
5000 LE Tilburg, The Netherlands
Tel.: 31 13 466 8216
E-mail: z.l.he@uvt.nl

Tony W. Tong *
Krannert School of Management
Purdue University
West Lafayette, IN 47907
Tel.: 1 765 494 4495
E-mail: tonytong@purdue.edu

Yuchen Zhang
Freeman School of Business
Tulane University
New Orleans, LA 70118
Tel.: 1 504 314 2467
E-mail: yzhang54@tulane.edu

Wenlong He
The Business School
University of International Business & Economics
Beijing, China 100029
E-mail: wenlong.he@uibe.edu.cn

December 14th, 2016

* Corresponding author.

1. Introduction

Patents provide highly detailed information on the innovation concerned, including technical descriptions, the assignee(s), the inventor(s), the time (application, grant, and expiry dates), the location (inventor and assignee addresses), the technology domains to which it belongs (technology classes), the scope of property rights (claims), the priority link (where a patent application has earlier been filed with foreign jurisdictions), and so forth. With such rich information, researchers can do more fine-grained analyses of innovation than what is possible with aggregate R&D expenditure data or self-reported innovation activities in survey data. Patent applications are filed by firms of various kinds, big and small, domestic and foreign, state-owned and non-state-owned, and across nearly all industries in a consistent manner. The large number of patents, often in the magnitude of millions, makes patent data the most comprehensive coverage of industrial innovation. Moreover, patents have been granted continuously by patent offices in different countries, and patent data often come in computer readable forms, making it possible to build wide and long panel datasets for sophisticated econometric analysis (Griliches, 1990).

While extremely valuable for innovation research, patent data do not come with firm identifiers that can be readily used to link to other sources of data, such as firms' financial and market data that are publicly available and collected on a regular basis. The absence of a unique identifier shared by assignees in patent data (USPTO, EPO, SIPO, etc.) and firms in widely used databases (Compustat, Orbis, CSMAR, etc.) is a major obstacle to harnessing the power of patent data for innovation research. This limitation is especially problematic for researchers who study the antecedents (e.g. corporate governance) or consequences (e.g. profitability) of innovation at the firm level. As warned by Hall, Jaffe, and Trajtenberg (2001: 24), any analysis based on patent data alone "would be self-contained, with all the limitations that implies". In order to take full advantage of the richness of patent data, a number of important projects have been carried out to match patents to the firms that own them. These include Hall et al. (2001), Belenzon, Berkovitz, and van Reenen (2007), Balasubramanian and Sivadasan (2010), Thoma, Torrisi, Gambardella, Guellec, Hall, and Harhoff (2010), among others.

Although the resultant datasets from these matching projects have greatly expanded the number of questions that researchers are able to analyze using patent data, they all focused on patent offices in developed economies, notably the USPTO and the EPO. Such a focus failed to give attention to the "Changing Face of Innovation" (WIPO, 2011)—the growing prominence of innovation and patenting activities in major emerging economies, such as China. China in particular has experienced a sustained, strong growth in patent filings over the past two decades. For instance, applications of invention patents, the type of patents that is most comparable across countries, to China's State Intellectual Property Office (SIPO) increased from 14,409 in 1992 to 526,412 in 2011, overtaking the U.S. to become the world's top patent filer (WIPO, 2012: 58), a position China has since been able to secure.

To meet the increasing interest in innovation activities in China and reduce potentially duplicative matching efforts, our research team has completed two projects that link patents filed with SIPO to two different sets of firms. In the first project, we matched SIPO patents to all firms listed on the Shanghai or Shenzhen Stock Exchange ("Main Board") and their subsidiaries during 1990-2010. Our first project has three key strengths: 1) we built comprehensive corporate family trees to dynamically record any changes of corporate affiliations and company names; 2) we implemented both computerized matching and careful

manual check to enhance efficiency and accuracy of our matching; and 3) we captured and coded, for each matched patent, the date of filing, date of publication, date of request for substantive examination, and dates of grant (if granted), withdrawal (if withdrawn), or refusal (if refused). A detailed description of this project is given in He, Tong, Zhang, and He (2016). An abridged project description along with matched patents are made publically available to the research community on a Google site "Chinese Patent Data Project" (https://sites.google.com/site/sipopdb). Despite its many strengths, this first project nonetheless had several limitations. For instance, there were only about 1,400 publicly listed firms in China for the matching period. These firms make up a small part of the Chinese economy and hence cannot account for the surge of patenting in China. Moreover, about two thirds of listed firms in China have the state as the ultimate controlling owner; privately- and collectively-owned firms as well as foreign-invested firms, which are an important driving force of innovation in China, are seriously underrepresented among listed firms.

To expand the scope of our patent matching effort as well as to address the abovementioned limitations of the first project, we conducted a second project to match SIPO patents to firms covered in China's Annual Survey of Industrial Enterprises (ASIE) 1998-2009.[1] Widely dubbed "the Census data" that are collected and maintained by China's National Bureau of Statistics (NBS), the ASIE covered from about 165,000 firms in 1998 to around 450,000 firms in 2009, representing the broad Chinese economy across different industries, different regions or provinces, and different types of ownership such as state-owned and non-state-owned firms, domestic and foreign firms, etc. Because of these attractive features, we believe a database linking SIPO patents to ASIE firms will be an invaluable contribution to research on Chinese patents and the growing innovation activities in China.

The remainder of this paper is structured as follows. We first provide a brief introduction to the ASIE database. Then we present our matching methodology and discuss several key decisions that we made and users of our matched patent dataset should be aware of. In the section that follows, we offer a brief summary of the matched patents and compare our matched results with several related matching efforts in prior studies. A final section concludes.

2. Annual Survey of Industrial Enterprises (ASIE)

The ASIE is a nationwide mandatory survey administered annually by NBS to firms in mining industries (6 two-digit industries from 06 to 11), manufacturing industries (30 two-digit industries from 13 to 37 and from 39-43), and utility industries (3 two-digit industries from 44 to 46). Therefore, the ASIE consists of mostly manufacturing firms. The ASIE database is also referred to as the Annual Industrial Survey Database (e.g. Chang and Wu, 2014; Zhang, Li, and Li, 2014), the Annual Census on Industrial Enterprises (Li, Zhou, and Zajac, 2009; Xu, Lu, and Gu, 2014), the China Industry Census (Huang, Jin, and Qian, 2013), the Annual Survey of Industrial Firms (Li, Lu, and Tao, 2016), or some other variants. The ASIE provides data for NBS to compute China's GDP (Cai and Liu, 2009), and the aggregated information is published in the official *China Statistical Yearbook* (Chang and Xu, 2008). It contains a large number of firms distributed across different industries and throughout the country's 31 provinces, autonomous regions, and province-equivalent municipalities (Beijing, Shanghai, Tianjin, and Chongqing). These firms account for around

[1] The database is called 工业企业数据库 in Chinese.

95% of total Chinese industrial output and 98% of total Chinese industrial exports (Eberhardt, Helmers, and Yu, 2011: 9; Tan and Peng, 2003: 1257). Table 1 reports the number of firms covered in our database by province and year. As the table shows, the total number of firms in our database for each year is consistent with the corresponding figure reported by NBS in the official *Yearbook* and that reported by a data vendor HuaMei Information.

Prior to 1998, NBS collected information only from state-owned firms and collectively-owned firms (Peng, Tan, and Tong, 2004; Xu et al., 2014: 528).[2] After 1998, the ASIE database covered all state-owned firms, and non-stated-owned firms—including privately-owned firms and foreign-invested firms—with annual sales of 5 million RMB (about 600-700 thousand U.S. dollars) or above (Chang and Xu, 2008; Yu, 2015). To ensure consistency in firm coverage, data prior to 1998 are often not used in recent research (e.g. Chang and Xu, 2008; Chang and Wu, 2014; Huang et al., 2013; Xu et al., 2014; Zhang et al., 2014). Year 2011 witnessed another major change: NBS raised the threshold for inclusion for all firms from 5 million to 20 million RMB of total sales (Xu et al., 2014: 528).[3] Since we only cover the ASIE data 1998-2009, our project is not affected by this change; in fact, it seems that post-2009 ASIE data have not been made available by NBS for academic research.

Besides a wide distribution across different provinces and different industries, the ASIE data exhibit large heterogeneity in terms of firm size and ownership. Despite the threshold of 5 million RMB of total sales,[4] most ASIE firms are in fact rather small. For instance, among 162,885 firms contained in the 2000 ASIE, 7,983 were classified by NBS as large firms (4.9%), 13,741 as medium-size firms (8.4%), and 141,161 as small firms (86.7%); among 336,768 firms contained in the 2007 ASIE, 2,910 were classified as large firms (0.9%), 33,596 as medium-size firms (10.0%), and 300,262 as small firms (89.2%).[5]

Furthermore, each ASIE firm is given a label of "registration type" based on 23 different ownership classifications (Jefferson et al., 2003: 112). Researchers often group them into broader categories for specific research purposes or to ease interpretation of results; in other cases researchers include only a few selected ownership classifications in their study or use information on six sources of a firm's registered capital (state capital, collective capital, legal person capital, individual capital, Hong Kong/Macao/Taiwan capital, and foreign capital) to code a number of ownership indicators (see Table 2). One important strength of the ASIE database is in fact the coverage of FDI firms: firms with investment from Hong Kong, Macao, and Taiwan (HMT) and foreign countries are more likely to be "above scale" and hence included in the ASIE (Huang et al., 2013: 871). The ASIE data show a clear trend of privatization: the proportion of state-owned and collectively-owned firms declined significantly from more than 70% in 1998 to less than 16% in 2005 as reported by Chang and Xu (2008: 502), as these firms were replaced by "reformed local firms" (private firms,

[2] In China, a collectively-owned firm is one that is owned collectively by its employees (typically organized as a workers' union) in cities or by local residents in rural areas (Peng et al., 2004; Xu et al., 2014: 529).

[3] This elevation of the total sales threshold, as expected, led to a significantly smaller firm coverage: the number of firms in the ASIE declined from 452,872 in 2010 to 325,609 in 2011 and 343,769 in 2012. See http://www.stats.gov.cn/tjsj/ndsj/2013/indexch.htm.

[4] Such firms are often referred to as "above scale" firms in the Chinese economy (see Cai and Liu, 1999), where "scale" means the threshold defined by NBS.

[5] See http://www.stats.gov.cn/tjsj/ndsj/2001c/m1303c.htm, http://www.stats.gov.cn/tjsj/ndsj/2008/indexch.htm. Note that the NBS definition of large- and medium-size enterprises (LMEs) among these "above scale" firms varied over time and across industries (Jefferson, Hu, Guan, and Yu, 2003: 91-92; Hu and Jefferson, 2009: 59).

shareholding firms, joint-stock firms, etc.) and to a lesser extent by HMT- and foreign-invested firms (Jefferson et al., 2003; Chang and Wu, 2014; Xia and Walker, 2015).

The ASIE database contains rich firm-level demographic, operational and financial information, including company name and address, year of founding, ownership details, 4-digit industry classifications,[6] three main products in order of relative importance, number of employees, total production output, total output accounted for by new products, total sales, exports, profits, total assets, intangible assets,[7] capital investment and depreciation, expenditure on R&D and advertising (available for selected years), total salary expenditure, employee training expenditure, etc. (Huang et al., 2013; Zhang et al., 2014). In total, more than 100 variables in the firms' main accounting statements (balance sheet, income statement, and statement of cash flows) are available (Yu, 2015: 949-950), making the ASIE the most comprehensive and detailed database of domestic and foreign firms operating in China. Moreover, each firm in the database has a unique firm identifier, known as the legal person code,[8] which allows researchers to assemble a panel dataset by linking together the firm's yearly observations (see e.g. Chang and Xu, 2008; Chang and Wu, 2014; Huang et al., 2013; Zhang, Li, Li, and Zhou, 2010; Zhang et al., 2014; Chen, Igami, Sawada, and Xiao, 2016).[9]

The information reported by ASIE firms to NBS should be fairly reliable for several reasons. First, firms are required by law to cooperate with surveys conducted by NBS (Chang and Xu, 2008), and they do not have clear incentives to misreport because the information reported to NBS cannot be used against them by other government agencies such as tax authorities (Cai and Liu, 2009).[10] Second, NBS has endeavored to maintain consistency and reliability in data collection (Zhou and Li, 2008). NBS has a Statistics Bureau in each county, with an office dedicated to collecting, maintaining, and reporting economic data. All the "above scale" firms have formally designated personnel who work with the Statistics Bureau to ensure that the reported data are reliable and consistent with NBS formats and definitions.

As a result, it is widely believed that the ASIE data are largely accurate and internally consistent for solid empirical work (Chow, 1993). Research based on the ASIE data has been published in leading journals in different fields, such as *Academy of Management Journal* (Zhang et al., 2014), *Administrative Science Quarterly* (Xu et al., 2014; Zhou, Gao, and Zhao,

[6] At this level, as pointed out by Huang et al. (2013: 871), researchers are able to distinguish "a firm producing leather shoes from a firm producing sneakers."

[7] Similar to accounting treatments in the U.S., intangible assets under Chinese accounting rules include patented and non-patented technology and know-how, brand names and trademarks, copyrights, various types of licensed rights and franchise rights, and goodwill (Huang et al., 2013: 881; Zhang et al., 2014: 706; Tian, 2007: 157).

[8] A firm's legal person code is the company ID created upon the legal creation and registration of the company (Fang, Lerner, and Wu, 2016: 17).

[9] However, it is not so unusual that a firm with exactly the same name and address acquires a different identifier due to ownership changes as a result of restructuring, joint ventures, mergers and acquisitions, etc. (Hu and Jefferson, 2009: 66; Chang and Xu, 2008: 502; Chang and Wu, 2014: 1109). Moreover, firms sometimes omit this field in their survey reports. Dealing with changing or missing firm identifiers requires care and discretion: some researchers assess the match of two firms' demographic information (name, industry classification, address, etc.) to determine whether they are indeed the same firm (Chang and Xu, 2008: 502; Fang et al., 2016: 17), while others regard different identifiers as representing different firms because ownership has changed so much as to require an identifier change (Chang and Wu, 2014: 1109).

[10] Article 33 of *Regulations on National Economic Census* states the following: "The use of materials regarding units and individuals collected in economic census shall be strictly limited to the purpose of economic census and shall not be used by any unit as the basis for imposing penalties on respondents of economic census."

2016), *American Economic Review* (Song, Storesletten, and Zilibotti, 2011), *American Journal of Sociology* (Walder, 1995), *Economic Journal* (Cai and Liu, 2009), *Journal of Economics & Management Strategy* (Li, Lu, and Tao, 2016), *Journal of International Business Studies* (Park, Li, and Tse, 2006), *Quarterly Journal of Economics* (Hsieh and Klenow, 2009), *Review of Economics and Statistics* (Hu, Jefferson, and Qian, 2005; Huang et al., 2013), and *Strategic Management Journal* (Chang and Wu, 2014).

Although the ASIE database provides the most comprehensive and accurate archival data on Chinese industrial firms, researchers find errors and omissions of various kinds when they use the data for academic research, possibly because of misreporting by firms and errors in data entry by NBS staff. To ensure the quality of the data, researchers often implement screens to remove potentially problematic observations (e.g. Cai and Liu, 2009; Yu, 2015: 950), drop firms that report "unrealistic" numbers (e.g. Chang and Xu, 2008: 504), or triangulate different but related information fields such as a firm's industry classification and the three self-reported main products (e.g. Huang et al., 2013: 871-872).

For innovation scholars, two fields contained in the ASIE database are particularly attractive: R&D expenditure and new product sales. However, both fields can be said to suffer from serious limitations. Other than the usual critique that R&D expenditure is too coarse an indicator of innovation, data on R&D expenditure are only available from 2005-2007 in the ASIE and contain many missing values (Nie, Jiang, and Yang, 2012). Similarly, information on new product sales is not consistently available across all firms and for all years. In addition, the definition of "new product" seems to vary across different years and across different studies.[11] Because of these limitations, relying on ASIE data alone limits the scope and type of questions that innovation scholars can address, as well as the level of granularity of the empirical evidence that can be produced. As a result, several studies have attempted to match ASIE firms to SIPO patents to better examine these firms' innovation activities (please see Section 4 below for more details). It is worth noting that researchers have also sought to amplify the value of the ASIE database by linking it with other databases, such as the Survey of Foreign-Invested Industrial Enterprises (e.g. Li et al., 2009), the Foreign Direct Investment Enterprise Database (e.g. Zhang et al., 2014), and the customs product-level trade data (e.g. Yu, 2015).

3. Matching Methodology

This matching project involves the following five steps.

[11] For instance, Zhou and Li (2008: 1122) explain that new products are "those new to the market" that either (1) adopt completely new scientific principles, technologies, or designs, or (2) are substantially improved in comparison with existing products in terms of performance and functionality, through significant changes in structure, materials, design, or manufacturing processes. By contrast, Chen et al. (2016: 20) indicate that the "'new product' definition is at the firm-year level and not always clear or uniform throughout the sample." Jefferson et al. (2003: 107) follow the Social, Population, and Technology Department of NBS to define new product as "new in relation to the reporting firm's prior product mix." In addition to varying definitions, the validity period of new products also differs across studies. For example, Zhou and Li (2008: 1122) indicate that a firm's new products are subject to the local government's certification, and that such certification is generally valid for up to 3 years. By contrast, Chen et al. (2016: 4) indicate that new products sales is the fraction of sales from the products that the firm introduced within one year.

3.1 Extracting patent data

We extract all patents from the SIPO patent database (DVD-ROMs, version April 2013), which covers all published patent applications since 1985 when SIPO started to accept patent applications. In total, the database contains 2,417,903 design patents, 3,063,153 invention patents, and 2,906,059 utility model patents. Because the ASIE database covers firms from 1998-2009, we remove patents with application dates outside the period of 1998-2009. In addition, because no individuals are included in the ASIE database and all firms covered are located in mainland China, we further remove the following two sets of patents before matching so as to improve computation efficiency: 1) patents assigned to individuals, and 2) patents assigned to firms with an address outside Greater China (mainland China, Hong Kong, Macao, and Taiwan). We do not exclude SIPO patents with an assignee address in Hong Kong, Macao, and Taiwan (HMT) mainly because as a result of increasing economic integration between the mainland and HMT, it is often difficult to discern whether an HMT firm is mainly based in the "mainland" or HMT. Such ambiguity is further exacerbated by the so-called "round-trip FDI" phenomenon whereby Chinese firms invest in the mainland under disguised foreign identities very often via Hong Kong (Huang, 2003: 37-14; Chang and Xu, 2008: 502).

A patent is considered to be assigned to an individual when it meets two conditions simultaneously: 1) its inventor(s) also appear in the assignee field, *and* 2) the assignee field does *not* contain any designators of corporate form. These designators include the following Chinese strings: 股份有限责任公司, 股份有限公司, 有限责任公司, 独立行政法人, 有限总公司, 有限分公司, 总公司, 分公司, 董事会, 集团, 有限公司, 有限责任, 株式会社, 公司, 股份, 企业, 工厂, and 厂. The *and* operator is used because firms and individual inventors may apply for patents jointly and as a result, the assignee field may contain both firm names and individual inventor names, and such patents should not be removed from the matching process. We use a long list of designators of corporate form to avoid incorrectly removing any patents that may be subsequently matched to firms in the ASIE database. Moreover, in case of 工厂 or 厂, we also require the length of the assignee to be shorter than four Chinese characters; this is because although some Chinese individuals may have 工厂 or 厂 in their names, Chinese individuals' names typically have only three or two characters.

Similar sufficient conditions are applied to identify and remove patents assigned to firms with an address outside Greater China: 1) the assignee country code is outside Greater China, *and* 2) the assignee is a firm, i.e. the assignee name contains one of the above designators of firm organization. It is crucial to point out that since assignee address and country code are available only for the first assignee (in case of multiple assignees), we do not remove any patents that have two or more assignees in the assignee field.

After removing the above three sets of patents that are irrelevant to our matching, we duplicate patents of multiple assignees into multiple records, one for each assignee. Such records have the same assignee field but different focal assignees. Put another way, we assume such a patent to be equally owned by all of the assignees, and match them one by one (also see He et al., 2016). To accelerate computerized matching, we format all patent records into nine tab-delimited text files, three for each type of patents, so that we run nine parallel threads at the same time. In total, there are 1,553,140 unique patent records to be matched with ASIE firms 1998-2009, including 468,972 unique design patents, 536,956 unique invention patents, and 547,212 unique utility model patents. Because there are important

differences among these three types of SIPO patents in terms of subject matter, examination procedure, patenting and maintenance costs, maximum possible protection period, etc. (see He et al., 2013), we match them separately to ASIE firms so as to generate three separate sets of firm-patent matches.

3.2 Preparing the list of ASIE firms

We face several challenges when compiling the list of ASIE firms for matching with patent data. First, missing firm identifiers and names are common in the 2009 ASIE data we had access to. Among 448,741 entries of the 2009 ASIE, 142,963 have firm identifier missing and 136,105 have firm name missing. For firms with missing identifiers, we search their names in earlier editions of the ASIE to find the same firm name, retrieve that firm's identifier, and use it for the year 2009. If the same firm name is found in multiple past years, the identifier for the year that is closest to 2009 will be used. A similar procedure is implemented to replace missing firm names wherever possible. Sporadic cases of missing firm identifiers and names in other years of the ASIE are dealt with in the same way.

Another issue is that there are obvious errors in some firm identifiers and names. For instance, we see strange firm identifiers such as 1 and 2 (too short to be a valid identifier), and]15174142 ("]" is an obvious error). We also see problematic firm names such as 鄂鄂州市隆昌合金钢有限责任公司 (the second 鄂 is obviously redundant perhaps as a data entry error), S 试笫星旆嵋□铣 (this firm name is completely messed up), and 6673896 (this may be a telephone number rather than firm name). Firm names like 海尔集团公司(本市) and 上海机床厂有限公司(本部) also pose problems because such content within brackets is likely to compromise matching accuracy. Spot checks by eyeballing firm names suggest that such problem cases only make up a very small minority in the vast ASIE database. Moreover, after some initial exploration, we realize that it would not be economical for us to spot and rectify such cases in a systematic and exhaustive manner. We therefore decide not to do anything with these obviously or potentially wrong firm identifiers and names. That said, the fact that we do "ever matching" instead of "contemporaneous matching" (more details in Section 3.4) should help minimize loss of true matches due to incorrect ASIE firm names: this is because as long as the firm has its correct name appearing once in the ASIE database 1998-2009, its entire patent portfolio for the period will be captured and matched. In other words, if a user finds that an ASIE firm should certainly have patents, but it does not appear in our matched output files, most likely this is because the firm has its name incorrectly spelt for *every* year in the ASIE database.

A third issue with ASIE firm names is that the same firm may appear in slightly different names in different years, sometimes with rather different operational and financial details, perhaps due to restructuring, mergers and acquisitions, or significant ownership changes. Moreover, a firm in the ASIE database may assume a different identifier even when its name and address remain unchanged at all.[12] After weighing substantial costs upfront against limited potential benefits, we opt not to consolidate ASIE firms before matching. Instead, matching is done for each unique firm identifier-name combination. To make an example, firm identifier 163578771 is linked to 海信集团 in some years but to 海信集团有限公司 in other years of the ASIE. Because we do not consolidate ASIE firm names and both 海信集团 and 海信集团有限公司 have the same stem name which is 海信 (see Section 3.3 for more

[12] Footnote 9 earlier provides more details.

details about stem names), the same set of patents would be matched to the firm identifier-name combination 163578771--海信集团 and to 163578771--海信集团有限公司 as well. The same applies to unique firm identifier-name combinations like 192352181--深圳市华为技术有限公司 and 192203821--深圳市华为技术有限公司, in which case two different firm identifiers are linked to exactly the same firm name. Admittedly, implementing this approach may lead to a large number of duplicate matches which we will assess in Section 4, but one advantage of this approach is that researchers using our matched patent data can consolidate ASIE firms in different ways according to their own specific research needs, and then the matched patents can be consolidated accordingly.

In the end, we arrive at 947,166 unique firm identifier-name combinations, among which 1,705 have missing identifiers. Then, we feed the patent data and the list of ASIE firms to our matching program for pre-processing routines (see Section 3.3) and then computerized matching (see Section 3.4), followed by manual check of computer outputted matches (see Section 3.5). We note that while we divide all patent data into nine different files to run nine parallel threads at the same time, a single file of ASIE firm names is fed into each thread.

3.3 Pre-processing patent assignee names and ASIE firm names

Before comparing patent assignee names and ASIE firm names, a set of pre-processing routines are implemented to clean and standardize names:

1) We trim all symbols and punctuation marks that are not letters, characters, or numbers. These include hyphen, parentheses, 《, apostrophe, comma, bar mark, etc. The content inside the parentheses, which often provides discriminating information, is kept. We make sure to remove both half-width and full-width symbols such as & and ＆, and both half-width and full-width punctuation marks such as ? and ？.[13]

2) We convert all full-width letters and numbers into half-width ones.

3) We convert Chinese numbers into Arabic numbers. Note that after implementing routine #2, all full width Arabic numbers (０, １, ２, …, ９) have already been converted into half-width ones (0, 1, 2, …, 9). Special care is needed to deal with intricate situations because firm or assignee names like 青岛四方机车车辆厂 contains one Chinese number 四, but it does not make much sense to convert 四 into 4 in this case, as 四 here does not mean anything strictly numerical.[14] Specifically, we sequentially implement the following steps:
    a. For three- or more digit numbers, convert 零/〇, 一, 二, …, 九 into 0, 1, 2, …, 9. For example, 中国人民解放军第三五零三工厂 will become 中国人民解放军第 3503 工厂.

[13] We compile a list of symbols and punctuation marks that could cause problems in matching based on the following sources: UTF-8 Chinese symbols, ASCII English symbols, Chinese punctuation marks from http://zh.wikipedia.org/wiki/标点符号, and ASCII English punctuation marks.

[14] With careful and systematic checks of both ASIE firm names and patent assignee names, we confirm that the complex form of Chinese numbers (壹, 贰, 叁, 肆, 伍, 陆, 柒, 捌, 玖, 拾) does not pose a threat to our matching.

b. For two-digit numbers, convert [一…九] 十 [一…九] into 11…99, and convert 一一, ..., 九九 into 11, …, 99.
c. For one-digit numbers, convert 第[一…十] into 第 [1…10].

4) We remove various designators of corporate form to obtain the so-called stem names. A set of such designators is the so-called stemming list, which includes 股份有限责任公司, 股份有限公司, 有限责任公司, 独立行政法人, 有限总公司, 有限分公司, 总公司, 分公司, 董事会, 集团, 有限公司, 有限责任, 株式会社 公司, 股份, 企业, 工厂, 厂.[15] Such designators appear in numerous names and should be removed before matching, because they provide little information to distinguish two names.

A few examples may help make sense of these pre-processing routines. Suppose we start with name ＴＣＬ－罗格朗国际电工(惠州)有限公司, it will become ＴＣＬ罗格朗国际电工惠州有限公司 with routine #1, become TCL 罗格朗国际电工惠州有限公司 with routine #2, no change with routine #3 as it does not contain any Chinese numbers, and then it will be stemmed into TCL 罗格朗国际电工惠州. To make another example, with these four pre-processing routines, both patent assignee name 天水二一三机床电器厂 and ASIE firm name 天水２１３机床电器厂 will be standardized and stemmed into 天水 213 机床电器, so that the correct patents will be matched. Likewise, both patent assignee name 蓝星成都六九一四电子设备厂 and ASIE firm name 蓝星成都６９１４电子设备厂 will become 蓝星成都 6914 电子设备, so that the corresponding true matches will be captured.

3.4 Matching ASIE firms to patent assignees

To choose an appropriate matching algorithm and design an efficient matching architecture, we need to make several important decisions, which are informed by careful tradeoffs between false positives (a name pair matched up by the computer program in fact refer to two different entities) and false negatives (a name pair not matched up by the computer program in fact refer to the same entity), and between matching accuracy and workload of manual check. It can be seen below that we approach each decision based on careful and iterative calibration, often with rounds of trial matching between 100,000 randomly selected patents and one year of ASIE data.

A matter of first order importance is the choice between approximate matching and exact matching. Approximate matching calculates a similarity score between two name strings and declares that a match is found when this score is above some threshold, whereas exact matching only identifies a pair of identical stem names. Approximate matching, however, comes at a price because it increases the probability of generating false positives which have to be eliminated by subsequent manual check. Eliminating such false matches by hand requires extensive time and manpower. Recall that our first matching project for listed firms generated 222,651 matches using approximate matching, and the subsequent manual check took about one month by three groups of research assistants working full time, plus two weeks by one of the authors to triangulate different manual check results (He et al., 2016).

[15] The same list is used to determine whether a patent is an individual patent or is assigned to a firm with an address outside Greater China (see Section 3.1).

Based on several trial runs, this second project, which involves a much larger number of firms, is expected to generate about one million matches, making the time and manpower required for manual check impractical. Another consideration is availability of computing power: while the average processing time is about 0.2-0.3 second when using exact matching to compare one patent assignee name against about 1 million entries in the master list of ASIE firm names, it is about 6-8 seconds when using approximate matching. Thus, to generate the estimated number of one million potential matches, exact matching would require a total of 83 computer hours, and approximate matching 2,499 hours, outstripping the computing resources available to us. We therefore adopt exact matching based on stem names for this project.

A key concern of exact matching is loss of true matches insofar as approximate matching can capture a pair of slightly different names that in fact refer to the same entity. Our assessment is that such loss of true matches should be very limited. Among 191,325 true matches in our first project, only 3,098 (1.6%) were captured due to the additional power of approximate matching, the remaining true matches would have been captured using exact matching or by applying the strict substring condition (see He et al., 2013). Assuming similar data structure for ASIE firms and considering the high cost of conducting approximate matching and manual check, we deem 1.6% loss of true matches acceptable. With two rounds of trial matching, we further confirm that the expected loss of true matches for ASIE firms due to not using approximate matching is around 1.6%.

Following our first project, we supplement exact matching by adding the condition of left-aligned strict substring to minimize loss of true matches. A case of left-aligned strict substring is where the stem ASIE firm name is a strict substring of the stem assignee name from the left. Table 3 shows that such name pairs are likely to be true matches, as the patent assignee is perhaps a division, factory, or branch office of the ASIE firm. However, this is not always the case, and thus such name pairs have to be manually checked to verify they are indeed true matches (see Section 3.5 for details). This extension to exact matching is particularly relevant for this project, because unlike the first project for which systematic corporate affiliation data were available for about 1,400 publically listed companies, it is not possible to collect and compile a list of subsidiaries for nearly one million ASIE firms. A natural question, however, is whether the strict substring condition should be allowed from both sides; in other words, whether the "left-aligned" restriction should be lifted to capture potentially more true matches. After multiple rounds of trail matching, we establish that without the "left-aligned" restriction, the number of computer generated matches that require manual check would increase by three to four times, yet a predominant majority of these additional matches are false matches. This is consistent with the convention of Chinese organization names where the largest entity comes first from the left, followed by the division, branch, or subsidiary name. We therefore opt to keep the "left-aligned" restriction.

Another important decision is the choice between "ever-match" and "contemporaneous match" (Balasubramanian and Sivadasan, 2010). While ever-match only requires the ASIE firm name and the patent assignee name to be matched, irrespective of in which year during 1998-2009 the firm appears in the ASIE database or the patent is filed with the SIPO, contemporaneous match adds the restriction that the ASIE firm year and the patent application year must coincide. We follow the ever-match logic primarily due to a common challenge in name matching: misspellings, name variations, and name changes. As noted above in Section 3.2, errors in ASIE firm names are not that rare. One key advantage of doing ever-match is that as long as the firm name is correctly recorded in one year of the

ASIE, all relevant patents will be found and matched. Ever-match is also necessary to account for name variations or name changes. For example, 深圳市中兴通讯股份有限公司 changes its name to 中兴通讯股份有限公司 in the ASIE database since 2004, yet in the SIPO database there are many patents under the new name before 2004 and many patents under the old name after 2004. Contemporaneous match would greatly undercount the company's patents in this case, as name changes in these two databases are not contemporaneous or synchronized. Moreover, as firms join or leave the ASIE database when they meet or fail to meet the threshold of 5 million RMB total sales, ever-match allows researchers to track a firm's patent portfolio over time even though the firm is absent in the ASIE during the intervening years. Ever-match also provides important flexibility post matching: should researchers prefer contemporaneous match, it could be easily achieved by imposing the restriction of *ASIE firm year = patent application year* to the outcomes of ever-match (we provide application year for each matched patent).

To summarize, after pre-processing ASIE firm names and patent assignee names, we conduct exact matching based on stem names, which is supplemented with strict substring matching to limit loss of potentially true matches, and we follow the ever-match logic for comprehensiveness and flexibility. With several final rounds of trial matching, we find that our matching logic is simple, clear, robust, and well-performing, doing what is designed to do. It is worth noting that the direction of matching, either from firms to assignees or from assignees to firms, does not make any substantive difference with exact matching, though such a direction matters greatly in terms of matching accuracy and efficiency with approximate matching in our first project (see He et al., 2016). The final matching program is implemented in Perl (V5.14.2) built for MSWin32-X64. The operating system is Windows 7 enterprise edition. This program requires Encode::CN, a module that can be found at "search.CPAN.org". By running nine threads at the same time, the computerized matching can be completed in less than four days.

3.5 Conducting manual check

The advantage of manual check is that, by using human cognition capability, we can incorporate additional information to more accurately identify a potentially matched name pair. However, manual check costs enormous efforts in searching for information and processing the information for a more confident decision. For this project, the automated matching program generates 1,155,649 matches in total, among which 1,010,471 are based on exact matching and 145,178 are based on left-aligned strict substring matching (see Table 4). Obviously, manually checking all the name pairs outputted by our computer program is not a viable approach. We therefore use a set of heuristics to first isolate those name pairs (one ASIE firm name and one patent assignee name) that require manual check.

More specifically, we make a heuristic-based program to go through all computer reported matches and mark whether a name pair needs manual check. First, when the assignee name in the name pair ends with 大学, 学院, 学校, 中学, or 小学, we deem it a case of false match and manual check not necessary. This is because such assignee names denote educational institutions, not profit-oriented industrial firms in the ASIE.

Second, when the ASIE firm name is fuzzy, showing ambiguous semantic meaning, we also mark such name pairs false matches and manual check not necessary. Table 5 gives all of the 36 such "fuzzy" names we identify from the whole list of ASIE firms. Matching for such names is meaningless and manual check cannot avail the situation.

Third, when the ASIE firm and the patent assignee have the same full name or when the firm's full name is a strict substring of the patent assignee's full name from the left,[16] we mark the name pair a case of true match and manual check not necessary. This is because under these two conditions, the chance of having false positives is virtually zero.

Then, for the remaining name pairs, we distinguish between cases of same stem names and cases of different stem names: while the former is based on exact matching, the latter is based on left-aligned strict substring matching. For cases of same stem names, we deem those with three or more characters (i.e. the length of stem names >=3) true matches and manual check not necessary, whereas all cases of stem name length shorter than 3 require manual check. Cases of different stem names generally require manual check: as explained above and shown in Table 3, the stem firm name being a strict substring of the stem assignee name from the left is no guarantee for true matches. Nevertheless, we isolate one type of cases which we can safely deem false matches without manual check: when the ASIE name has 大学, 学院, 学校, 中学, or 小学 anywhere in the name. The logic is that such ASIE names must refer to factories or companies affiliated to an educational institution, which has been common in China; therefore the stem names must fully match if they refer to the same entity. One example is the name pair of ASIE firm 上海交通大学附属工厂 (stem name 上海交通大学附属) vs. patent assignee 上海交通大学附属第一人民医院 (stem name 上海交通大学附属第一人民医院): they are obviously two different units under 上海交通大学.

As a result of the above procedures, we isolate 122,897 matches from the automated matching program that require manual check. These 122,897 matches are collapsed into 9,703 unique name pairs for manual check, which is carried out based on the following protocol or steps:

1. The first three authors did manual check for the same set of randomly selected 100 name pairs. By comparing the three sets of results, we discussed the potential reasons for differences and developed a manual check guideline, summarizing three rules:
   a. Accept as true match when a) the ASIE firm and the patent assignee are the same entity, b) one is part of the other, or c) one is a predecessor or successor of the other.
   b. Accept as true match when no information is available to clearly separate the two entities, e.g. when information is hardly available for one of the two names.
   c. If the ASIE firm and the patent assignee are located in the same province or city whose name appears in the firm and assignee names, and the two names are very similar (suggesting that one may be part of the other, but there is no unambiguous information indicating so), we accept the name pair as true match.

2. We hired six senior undergraduate students from a top university in China as research assistants. We gave them a workshop to explain the manual check procedure. Each student was subsequently given a same set of 50 name pairs for manual check. A meeting was then held to discuss and sort out any differences in their manual check outcomes.

3. These six RAs were divided into two teams to do manual check for all 9,703 unique name pairs by following the above protocol. Within each team, each RA dealt with around

[16] Note that in automated matching described earlier, this left-aligned strict substring rule is applied to stem names.

3,200 name pairs. With this procedure, two manual check results were independently obtained for each name pair. These two sets of manual check had a Cohen's Kappa of 0.77, indicating sufficient agreement between the coders.

4. The first author did the final check. If a name pair received unanimous "yes—this is a case of true match" or unanimous "no—this is not a case of true match", it would be accepted as it was. Otherwise, additional search was conducted to make a final decision.

By the end of this process, out of 122,897 computer generated matches that require manual check, we could confirm that 90,860 are true matches.

4. Summary of matched patents and comparison with prior studies

Our computer matching program obtains 1,155,649 matches for firms in the ASIE database 1998-2009, among which 1,113,588 are "true" matches. The distribution among the three different types of patents can be found in Table 6. Moreover, as explained earlier, significant duplicate matches are expected since we do not consolidate ASIE firm names. Table 6 shows that among 1,113,588 true matches, 849,647 patents are uniquely matched. Recall that a total of 1,553,140 unique patent records are fed into our matching program (see Section 3.1)—this means that around 55% of the patents in the pool are matched to ASIE firms, confirming the broad coverage of Chinese firms in the ASIE database. The table also indicates that duplicate matches account for less than a quarter of the matched patents. As suggested earlier, users of our dataset may choose different ways to remove duplicate matches based on their preferred approach to consolidating ASIE firm names.

We are aware of four prior studies that combined SIPO patent data with ASIE data in their research. In one study, Hu and Jefferson (2009) use firms' self-reported number of patent applications to examine what is behind China's patent explosion, based on a sample of approximately 20,000 large and medium-sized enterprises (LMEs) drawn from the ASIE database. However, self-reported patent numbers are likely subject to problems of unreliable recalls or "retrospective response bias" (Hall et al., 2007: 9).[17] Moreover, their study makes no distinction among the three different types of Chinese patents, a point also made by Dang and Motohashi (2015).

In three other studies, researchers conducted the matching of SIPO patents to ASIE firms. In the first study, Eberhardt et al. (2011) match SIPO patents contained in the PATSTAT database to ASIE firms 1999-2006. Their matching is "indirect" in that they use BvD's Oriana database to bridge firm names that appear in the ASIE (Chinese characters only) and those appearing in PATSTAT (pinyin transcription, English translation, or a mix of both). This indirect approach raises several concerns. First, BvD has very limited coverage of Chinese firms. In fact, the authors acknowledge that "The Oriana version available to us contains firm-level data for about 23,000 Chinese firms for the period 2000-2005" (p. 10), which is a small subset of the ASIE data that contain about 200,000 firms each year during

[17] Substantial and non-random errors in self-reported patent counts are found in similar survey data. For example, the UK Community Innovation Survey 3 (2001) contains a question about the number of patents a respondent firm applied for during the period of 1998 to 2000, but Eberhardt et al. (2011: 9) find that "cross-checking firms' responses to this question with their actual patent holdings indicates that only about 30% of firms that report to have applied for a patent actually did so."

this period. Second, while PATSTAT contains most of the SIPO invention and utility model patents, its coverage of SIPO design patents is very poor (He et al., 2016). Third, names in PATSTAT and Oriana might differ substantially depending on whether they are transcribed using pinyin or (partly) translated into English; this added complication might greatly reduce matching accuracy and efficiency.

In a second study, Dang and Motohashi (2015) match ASIE firms with SIPO patents. Their matching of all the observations in the ASIE data (1998-2008) with patent data by names of companies leads to 126,386 invention patent applications from 12,208 firms. The authors report a table showing the distribution of matched patents by year, but do not provide details about how their matching approach is designed and implemented.

In a third, most recent study, Chen et al. (2016) match ASIE data with SIPO patent data, but no explanation whatsoever is given regarding how the two types of data are linked together.

Table 7 provides a comparison of our number of matched patents with those reported in the three abovementioned studies. As the last two columns in the table show, the number of unique patents matched in our project is much larger than the corresponding numbers reported in the three previous matching efforts. We see three major reasons for the possibly superior performance of our matching approach. First, in contrast to Eberhardt et al. (2011), we match firm names in the ASIE database to assignee names in the SIPO database directly, both of which are in Chinese characters. Second, we implement ever matching such that as long as a firm appears in one year of the ASIE data, all of its patents during the period will be captured and matched. This is different from Dang and Motohashi (2015) and Chen et al. (2016) who actually implement contemporaneous matching by first isolating a balanced panel of firms from the ASIE database. Finally, we design and implement a systematic, iterative approach that includes multiple calibration procedures and manual check.

5. Conclusion

In this project, we carefully designed a data parsing and pre-processing stage to clean and stem firm and assignee names, selected a matching algorithm that fits with our data and maintains a balance between matching accuracy and workload of manual check, and implemented a systematic manual check process to filter out false positives generated from computerized matching. After numerous hours of intensive efforts where almost every minor detail was a major decision, the project finally reached fruition.

To encourage future research on the new innovation landscape in China, we are making the matched dataset publicly available to the research community on the Google site “Chinese Patent Data Project.” The shared dataset comes in three different Excel files, each for a different type of patents. In each file, Columns A-C record the ASIE firm identifier, firm name, and stem firm name, respectively; Columns D-I record the type of the patent, a serial number we created to mark the sequence in which the patent appears in the SIPO patent database, year of patent filing, assignee field, focal assignee name, and stem assignee name, respectively; Columns J-K mark whether manual check is necessary for the name pair and whether it is a case of true match; and the remaining columns give additional information on the matched patent.

By sharing our matched patent dataset with the research community, we hope to reduce

duplicative matching effort and motivate more research on the fast growing innovation activities taking place in China. We hope that researchers will use the dataset in novel and productive ways and that they will help to improve the dataset over time to make it a valuable community resource.

6. References

Balasubramanian N, Sivadasan J (2010) NBER Patent Data-BR Bridge: User Guide and Technical Documentation. CES Discussion Papers CES 10-36, U.S. Census Bureau, Washington, DC.

Belenzon S, Berkovitz T, van Reenen J (2007) AmaPat – Innovation, Ownership and Financials for European Firms. Available at: http://ssrn.com/abstract=1022044 and the related PowerPoint presentation at http://papers.ssrn.com/paper=1027909

Cai H, Liu Q (2009) Competition and corporate tax avoidance: Evidence from Chinese industrial firms. *Economic Journal*, 119(537): 764-795.

Chang SJ, Wu B (2014) Institutional barriers and industry dynamics. *Strategic Management Journal*, 35(8): 1103-1123.

Chang SJ, Xu D (2008) Spillovers and competition among foreign and local firms in China. *Strategic Management Journal*, 29(5): 495-518.

Chen Y, Igami M, Sawada M, Xiao M (2016) Privatization and innovation: Productivity, new products, and patents in China. Working Paper, Peking University, Yale University, and University of Arizona.

Chow GC (1993) Capital formation and economic growth in China. *Quarterly Journal of Economics*, 108(3): 809-842.

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

Eberhardt M, Helmers C, Yu Z (2011) Is the dragon learning to fly? An analysis of the Chinese patent explosion. CASE Working Paper, WPS/2011-15.

Fang L, Lerner J, Wu C (2016) Intellectual property rights protection, ownership, and innovation: Evidence from China. Cambridge, MA: NBER Working Paper No. 22685.

Griliches Z (1990) Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4): 1661-1707.

Hall BH, Jaffe AB, Trajtenberg M (2001) The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools. Cambridge, MA: NBER Working Paper No. 8498.

Hall BH, Thoma, G, Torrisi S (2007) The market value of patents and R&D: Evidence from European firms. Cambridge, MA: NBER Working Paper No. 13426.

He ZL, Tong TW, Zhang Y, He W (2016) Constructing a Chinese patent database of listed firms in China: Descriptions, lessons, and insights. *Journal of Economics and Management Strategy*. Forthcoming.

Hsieh C, Klenow P (2009) Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics*, 124: 1403-1448.

Hu AG, Jefferson GH (2009) A great wall of patents: What is behind China's recent patent explosion? *Journal of Development Economics*, 90(1): 57-68.

Hu AG, Jefferson, GH, Qian J (2005) R&D and technology transfer: Firm-level evidence from Chinese industry. *Review of Economics and Statistics*, 87(4): 780-786.

Huang Y (2003) *Selling China: Foreign Direct Investment during the Reform Era*. Cambridge University Press: Cambridge, UK.

Huang Y, Jin L, Qian Y (2013) Does ethnicity pay? Evidence from overseas Chinese FDI in China. *Review of Economics and Statistics*, 95(3): 868-883.

Jefferson G, Hu AG, Guan X, Yu X (2003) Ownership, performance, and innovation in China's large- and medium-size industrial enterprise sector. *China Economic Review*, 14(1): 89-113.

Li H, Lu Y, Tao Z (2016) Vertical integration and firm productivity. *Journal of Economics & Management Strategy*. Forthcoming.

Li J, Zhou C, Zajac EJ (2009) Control, collaboration, and productivity in international joint ventures: Theory and evidence. *Strategic Management Journal*, 30(8): 865-884.

Nie H, Jiang T, Yang R (2012). A review and reflection on the use and abuse of Chinese industrial enterprise database. *Journal of World Economy*, 2012(5): 142-158 (In Chinese).

Park SH, Li S, Tse DK (2006) Market liberalization and firm performance during China's economic transition. *Journal of International Business Studies*, 37(1): 127-147.

Peng MW, Tan J, Tong TW (2004) Ownership types and strategic groups in an emerging economy. *Journal of Management Studies*, 41(7): 1105-1129.

Song Z, Storesletten K, Zilibotti F (2011) Growing like China. *American Economic Review*, 101(1): 202-241.

Tan J, Peng MW (2003) Organizational slack and firm performance during economic transitions: Two studies from an emerging economy. *Strategic Management Journal*, 24(13): 1249-1263.

Thoma G, Torrisi S, Gambardella A, Guellec D, Hall BH, Harhoff D (2010) Harmonizing and Combining Large Datasets: An Application to Firm-level Patent and Accounting Data. Cambridge, MA: NBER Working Paper No. 15851.

Tian X (2007) Accounting for sources of FDI technology spillovers: Evidence from China.

Journal of International Business Studies, 38(1): 147-159.

Walder A (1995) Local governments as industrial firms: An organizational analysis of China's transition economy. *American Journal of Sociology*, 101(2): 263-301.

WIPO (2011) *World Intellectual Property Report: The Changing Face of Innovation*. World Intellectual Property Organization, Geneva, Switzerland.

WIPO (2012) *World Intellectual Property Indicators*. World Intellectual Property Organization, Geneva, Switzerland.

Xia F, Walker G (2015) How much does owner type matter for firm performance? Manufacturing firms in China 1998-2007. *Strategic Management Journal*, 36(4): 576-585.

Xu D, Lu JW, Gu Q (2014) Organizational forms and multi-population dynamics: Economic transition in China. *Administrative Science Quarterly*, 59(3): 517-547.

Yu M (2015) Processing trade, tariff reductions and firm productivity: Evidence from Chinese firms. *Economic Journal*, 125(585): 943-988.

Zhang Y, Li H, Li Y, Zhou L (2010) FDI spillovers in an emerging market: The role of foreign firms' country origin diversity and domestic firms' absorptive capacity. *Strategic Management Journal*, 31(9): 969-989.

Zhang Y, Li Y, Li H (2014) FDI spillovers over time in an emerging market: The roles of entry tenure and barriers to imitation. *Academy of Management Journal*, 57(3): 698-722.

Zhou C, Li J (2008) Product innovation in emerging market-based international joint ventures: An organizational ecology perspective. *Journal of International Business Studies*, 39(7): 1114-1132.

Zhou KZ, Gao GY, Zhao H (2016) State ownership and firm innovation in China: An integrated view of institutional and efficiency logics. *Administrative Science Quarterly*. Forthcoming.

Table 1. Distribution of ASIE firms across 31 provinces, autonomous regions, and province-equivalent municipalities, 1998-2009

Province	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Anhui	3,820	3,782	3,681	3,674	3,920	4,159	4,793	5,278	6,523	8,111	11,112	14,516
Beijing	4,502	5,240	4,575	4,327	4,557	4,024	6,906	6,297	6,400	6,397	6,886	6,934
Fujian	6,103	5,548	6,006	6,554	7,474	9,211	11,953	12,396	13,755	15,178	16,898	18,743
Gansu	1,654	2,253	2,859	3,097	3,217	2,894	2,022	1,733	1,733	1,841	1,807	1,993
Guangdong	17,977	18,881	19,695	20,652	22,620	24,519	34,738	35,157	37,494	42,260	51,134	53,422
Guangxi	3,365	3,146	3,159	3,059	2,913	2,873	3,751	3,686	4,049	4,408	5,089	5,841
Guizhou	2,051	2,121	2,088	1,923	2,069	2,123	2,546	2,584	2,594	2,295	2,501	2,807
Hainan	640	578	596	573	602	620	634	616	595	488	521	500
Hebei	7,524	7,337	7,164	7,511	7,536	7,816	9,345	9,938	10,633	10,870	12,192	13,509
Henan	10,445	9,913	9,924	9,644	9,663	9,089	11,741	10,867	11,895	13,510	17,825	18,754
Heilongjiang	3,558	2,995	2,716	2,504	2,635	2,612	3,345	2,887	2,956	3,172	4,296	4,496
Hubei	7,398	6,871	6,281	6,146	6,176	6,272	6,366	6,814	7,546	8,995	11,759	14,214
Hunan	4,557	4,797	4,808	4,779	5,439	5,959	7,610	8,022	8,999	10,201	11,345	13,509
Jilin	2,845	2,837	2,768	2,606	2,622	2,343	3,451	2,774	3,249	3,984	5,151	6,133
Jiangsu	17,997	18,004	18,313	19,610	21,467	23,856	40,899	32,224	36,319	41,841	63,610	63,380
Jiangxi	3,951	3,737	3,556	3,105	3,085	3,054	4,263	4,403	5,333	6,028	6,750	7,712
Liaoning	6,250	5,806	6,018	5,693	6,018	6,844	11,458	11,509	14,754	16,556	21,124	26,276
Inner Mongolia	1,368	1,280	1,262	1,195	1,320	1,531	2,284	2,448	3,074	3,363	3,691	4,639
Ningxia	539	521	435	424	408	437	666	685	761	745	876	1,007
Qinghai	570	555	441	378	396	398	478	406	437	473	470	552
Shandong	11,443	11,432	11,721	12,149	13,508	16,226	23,916	27,540	31,936	36,145	41,927	46,671
Shanxi	3,934	3,349	3,280	3,021	3,460	3,610	5,067	4,441	4,671	4,472	4,296	4,049
Shaanxi	2,683	2,587	2,551	2,329	2,464	2,489	3,117	2,998	3,372	3,373	3,702	4,494
Shanghai	9,401	9,340	8,588	9,745	10,094	11,126	15,766	14,806	14,403	15,099	18,291	18,043
Sichuan	4,982	4,542	4,399	4,477	4,907	5,434	7,454	7,958	8,995	10,709	13,258	13,461
Tianjin	5,437	5,245	5,465	5,598	5,376	5,381	6,466	6,145	6,302	6,361	7,658	8,639
Xizang	340	328	361	363	345	324	187	195	202	98	74	90
Xinjiang	1,821	1,627	1,456	1,326	1,266	1,256	1,446	1,445	1,481	1,575	1,807	2,022
Yunnan	2,514	2,131	2,122	1,988	2,070	1,992	2,398	2,362	2,603	2,699	2,994	3,547
Zhejiang	13,450	13,274	14,552	18,549	21,869	25,508	41,369	40,277	45,688	51,604	57,739	62,271
Chongqing	1,997	1,976	2,042	2,032	2,061	2,242	2,657	2,943	3,208	3,916	5,987	6,481
Province unknown	2	0	1	0	0	0	0	1	0	0	47	36
Our total	165,118	162,033	162,883	169,031	181,557	196,222	279,092	271,835	301,960	336,767	412,817	448,741
HMI total	165,119	162,034	162,885	169,031	181,557	196,222	279,092	271,835	301,961	336,768	412,000	434,000
NBS total	165,080	162,033	162,885	171,256	181,557	196,222	276,474	271,835	301,961	336,768	426,113	434,364

Notes: HMI (HuaMei Information) total can found at http://www.allmyinfo.com/data/zggyqysjk.asp, NBS total can be found at http://www.stats.gov.cn/tjsj/ndsj.

Table 2. Aggregation of ownership classifications in selected studies using the ASIE data

Article	Journal	Aggregation of ownership classifications	Remarks
Jefferson et al. (2003)	*China Economic Review*	SOEs, COEs, private firms, shareholding firms, HMT firms, non-HMT foreign firms, and other domestic firms	SOEs is short for state-owned enterprises, COEs is short for collectively-owned enterprises, HMT refers to Hong Kong, Macao, and Taiwan.
Chang & Xu (2008)	*Strategic Management Journal*	HMT firms, non-HMT foreign firms, conventional local firms, and reformed local firms	"conventional local firms" refer to SOEs and COEs, while "reformed local firms" include private firms, shareholding firms, and limited liability firms, which comprise the group of modernized and restructured Chinese companies.
Zhou & Li (2008)	*Journal of International Business Studies*	International Joint Ventures (IJVs)	This study only includes equity joint ventures formed by local (state or non-state) and foreign (HMT or non-HMT) companies and examines how ownership balance, state ownership, FDI legitimation, regional agglomeration, etc. affect IJVs' product innovation.
Li et al. (2009)	*Strategic Management Journal*	International Joint Ventures (IJVs)	This study only includes equity joint ventures formed by local (state or non-state) and foreign (HMT or non-HMT) companies, with the restriction that foreign equity shares must fall in the range of 5% to 95% of total equity.
Hsieh & Klenow (2009)	*Quarterly Journal of Economics*	SOEs, COEs, domestic private firms, and foreign firms	Domestic private firms include all other domestic firms that are not SOEs or COEs, and foreign firms include both HMT firms and non-HMT foreign firms.
Cai & Liu (2009)	*Economic Journal*	SOEs, COEs, private firms, HMT firms, non-HMT foreign firms, and mixed-ownership firms	Mixed-ownership firms are mainly joint stock companies, including publicly listed companies in China, which are not treated by NBS separately.
Hu & Jefferson (2009)	*Journal of Development Economics*	SOEs, COEs, private firms, limited liability firm, joint-stock firms, shareholding firms, HMT firms, and non-HMT foreign firms	
Zhang et al. (2010) Zhang et al. (2014)	*Strategic Management Journal, Academy of Management Journal*	Local firms, foreign firms	These two categories are defined by 100% ownership, i.e., local firms are 100% owned by domestic investors (state or non-state), and foreign firms are 100% owned by HMT or foreign investors.
Song et al. (2011)	*American Economic Review*	SOEs, COEs, DPEs, and FEs	DPEs means domestic private enterprises, and FEs refers to foreign-invested firms including both HMT firms and non-HMT foreign firms.
Huang et al. (2013)	*Review of Economics and Statistics*	HMT firms, non-HMT foreign firms	This study compares performance of HMT-invested firms and non-HMT foreign-invested firms China. Other types of firms are excluded from the analysis.
Xu et al. (2014)	*Administrative Science Quarterly*	SOEs, COEs, and POEs	POEs are defined broadly by including private firms, limited liability firms, and shareholding firms, not just "private firms" that in China specifically refer to firms owned by private individuals. SOEs, COEs, and POEs are seen as the old, transitional, and new organizational forms respectively in the Chinese economy.
Yu (2015)	*Economic Journal*	SOEs indicator and Foreign ownership indicator	SOEs indicator = 1 if a firm has any investment from the government, and 0 otherwise; Foreign ownership indicator =1 if a firm has any investment from HMT or foreign countries, and 0 otherwise.
Xia & Walker (2015)	*Strategic Management Journal*	SOEs, non-SOE local firms, and foreign firms	These three categories are defined by 100% ownership: SOEs are 100% owned by the state, non-SOE local firms refer to Chinese firms that have zero state ownership, and foreign firms are owned 100% by investors from HMT or foreign countries.

Notes: Disaggregated ownership classifications used by NBS can be found in Jefferson et al. (2003: 112) or at http://www.allmyinfo.com/data/zggyqysjk.asp.

Table 3. Examples of the left-aligned strict substring condition

ASIE firm name	Stem firm name	Patent assignee name	Stem assignee name	True match?
贵州黄果树烟草集团公司	贵州黄果树烟草	贵州黄果树烟草集团有限责任公司贵阳烟叶购销分公司原料分厂	贵州黄果树烟草贵阳烟叶购销原料分	Yes
鞍山钢铁集团公司	鞍山钢铁	鞍山钢铁集团公司水泥厂	鞍山钢铁水泥	Yes
长飞光纤光缆有限公司	长飞光纤光缆	长飞光纤光缆（上海）有限公司	长飞光纤光缆上海	Yes
上海创开无框阳台有限公司	上海创开无框阳台	上海创开无框阳台窗有限公司	上海创开无框阳台窗	Yes
上海宝钢集团公司	上海宝钢	上海宝钢建筑工程设计研究院	上海宝钢建筑工程设计研究院	Yes
北京市北郊冷饮食品厂	北京市北郊冷饮食品	北京市北郊冷饮食品三厂	北京市北郊冷饮食品三	No
天津市自动化仪表厂	天津市自动化仪表	天津市自动化仪表七厂	天津市自动化仪表七	No
洛阳市工程机械厂	洛阳市工程机械	洛阳市工程机械设计所	洛阳市工程机械设计所	No

Table 4. Breakdown of results from automated matching program

Design patents								Invention patents								Utility model patents							
398,483								332,682								424,484							
Exact matching				Left-aligned strict substring matching				Exact matching				Left-aligned strict substring matching				Exact matching				Left-aligned strict substring matching			
356,121				42,362				291,449				41,233				362,901				61,583			
Manual check required		Manual check not required		Manual check required		Manual check not required		Manual check required		Manual check not required		Manual check required		Manual check not required		Manual check required		Manual check not required		Manual check required		Manual check not required	
575		355,546		38,621		3,741		578		290,871		32,195		9,038		686		362,215		50,242		11,341	
True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches	True matches	False matches
551	24	355,527	19	29,022	9,599	2,150	1,591	532	46	290,094	777	22,983	9,212	3,659	5,379	605	81	361,833	382	37,167	13,075	9,465	1,876

Notes:
Total number of computer generated matches based on exact matching = 356,121 + 291,449 + 362,901 = 1,010,471.
Total number of computer generated matches based on left-aligned strict substring matching = 42,362 + 41,233 + 61,583 = 145,178.
Total number of matches that require manual check = 575 + 38,621 + 578 + 32,195 + 686 + 50,242 = 122,897, among which 90,860 are true matches and 32,037 are false matches.
Total number of matches that do not require manual check = 355,546 + 3,741 + 290,871 + 9,038 + 362,215 + 11,341 = 1,032,752, among which 1,022,728 are true matches and 10,024 are false matches.

Table 5. List of “fuzzy” ASIE firm names

1	印刷厂
2	建材厂
3	机械厂
4	林化厂
5	电机厂
6	电缆厂
7	食品厂
8	无线电厂
9	水电公司
10	油脂集团
11	热电公司
12	物流公司
13	电业公司
14	电力公司
15	黄金公司
16	天然食品厂
17	建筑材料厂
18	水电总公司
19	电力总公司
20	石油化工厂
21	矿业总公司
22	第三化工厂
23	有色金属公司
24	汽车零部件厂
25	油脂有限公司
26	电力工业公司
27	电力集团公司
28	电装有限公司
29	造纸有限公司
30	医药集团有限公司
31	电力有限责任公司
32	食品集团有限责任公司
33	2
34	鼎盛
35	微生物研究所
36	陶瓷有限责任公司

Table 6. Assessing the extent of duplicate matches

	Computer generated matches (a)	Unique patents matched (b)	% (b/a)	True matches after manual check (c)	Unique patents matched (d)	% (d/c)
Design patents	398,483	300,956	75.5	387,250	291,578	75.3
Invention patents	332,682	265,338	79.8	317,268	253,628	79.9
Utility model patents	424,484	316,303	74.5	409,070	304,441	74.4
Total	1,155,649	882,597	76.4	1,113,588	849,647	76.3

Table 7. Comparing our number of matched patents with that in prior studies

Authors	ASIE period	Number of firms	Distinguish different types of patents?	Number of patents matched	Our number of unique patents matched during the same period
Eberhardt et al. (2011) [a]	1999-2006	about 590,000	Only invention patents are matched	44,344 (invention patents)	95,902 [b] (invention patents)
Dang & Motohashi (2015)	1998-2008	12,208 (panel data)	Only invention patents are matched	126,386 (invention patents)	188,773 (invention patents)
Chen et al. (2016)	1998-2007	11,631 (panel data)	Yes	50,013 [c] (three types of patents in total)	484,359 (three types of patents in total)

Notes:

a. Eberhardt et al. (2011) match SIPO patents contained in the PATSTAT database to ASIE firms indirectly via Oriana. They acknowledge that "Oriana only contains a subset of the firms contained in the census" (p. 11) and "The Oriana version available to us contains firm-level data for about 23,000 Chinese firms for the period 2000-2005" (p.10).

b. Despite using ASIE data 1999-2006, Eberhardt et al. (2011) count invention patents applied for between 1985 and 2006 (see the Appendix therein). We report here the number of matched invention patents filed between 1999 and 2006, making a conservative comparison with theirs.

c. This is obtained by multiplying the number of observations (116,310) and mean number of patent applications (0.43, counting all three types of patents).