NBER PATENT DATA-BR BRIDGE:

USER GUIDE AND TECHNICAL DOCUMENTATION

by

Natarajan Balasubramanian *
Syracuse University

and

Jagadeesh Sivadasan *
University of Michigan

CES 10-36 October, 2010

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.ces.census.gov or contact Cheryl Grim, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K130B, 4600 Silver Hill Road, Washington, DC 20233, CES.Papers.List@census.gov.

Abstract

This note provides details about the construction of the NBER Patent Data-BR concordance, and is intended for researchers planning to use this concordance. In addition to describing the matching process used to construct the concordance, this note provides a discussion of the benefits and limitations of this concordance.

* Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. We thank Clint Carter, Kevin McKinney, and Arnie Reznek for prompt processing of our disclosure and data access requests. We also thank Cheryl Grim for a careful reading of an earlier draft of this note. Support for this research at the UCLA CCRDC from NSF (ITR-0427889) is also gratefully acknowledged. Unless stated otherwise, all company names and identifiers used in this document are fictional. They are for illustration purposes only.

1. Introduction

Technological innovation is generally considered to be among the most important sources of economic progress in general, and productivity growth in particular. However, measuring innovation has been a challenge for many reasons beginning with the lack of a commonly agreed definition of what constitutes an innovation. One of the most common approaches has been to proxy innovation with the input into the innovative process, typically the amount of research and development (R&D) expenditure. Another approach has been to use the outputs of the innovative process such as patents and new product introductions. However, for many years, output-based approaches were constrained by the lack of any large-scale data. A giant leap forward in measuring innovation was taken when a team of researchers led by Bronwyn Hall, Adam Jaffe and Manuel Trajtenberg (2001) constructed the NBER Patent Dataset. This immense exercise, lasting a number of years, compiled front-page data on *all* utility patents granted by the United States patent and trademark office (USPTO) between 1963 and 1999. These data have now been used in dozens, perhaps hundreds, of studies on innovation. In this project, we extend the efforts of Hall, Jaffe and Trajtenberg by building a Bridge that would allow researchers to link patent data from the NBER Patent Dataset to firm level data available at the US Census Bureau.

This project was undertaken to meet two unfulfilled requirements of researchers, particularly at the US Census Bureau. First, the NBER Patent Dataset offered researchers an opportunity to link firm data in Compustat to patent data; however, there was no such link available for researchers interested in linking Census data to patent data. The second need arose

out of one of the limitations in the NBER Patent Dataset.[2] The existing link between Compustat and the patent data is based on the 1989 Compustat universe of firms. Hence, the match reflects all corporate relations (e.g. some patent assignees may be subsidiaries of bigger corporations) as at 1989.[3] This implies that ownership changes prior to or subsequent to 1989 would not be reflected accurately in the match. Given that the US Census Bureau spends large sums of money in tracking ownership changes for other purposes, it seemed natural to use the results of those efforts to extend the coverage offered by the NBER Patent Dataset. This project was a step in that direction.

The end-product of this project is the NBER Patent Data-BR Bridge (or the Bridge). The Bridge is a concordance between the assignees in the NBER Patent Dataset and firms in the Business Register (BR) of the Census Bureau.[4] The objective of the concordance is to facilitate the linking of patent data with firm data collected by the US Census Bureau. The concordance was developed using name matching algorithms that used the names of assignees and firms, along with geographic variables such as city and state to generate matches between assignees and firms. The concordance is available for every year that the assignee is matched to a firm in the BR. We hope that this concordance will significantly extend the number of questions that

[2] As an aside, this need was suggested to one of us by Manuel Trajtenberg during his visit to UCLA around 2004. He also suggested a matching algorithm (that used the fact that the probability of two assignees with the same name being in the same SMSA was very low) close to the one eventually used by us. We thank him for the suggestions.

[3] As one of the referees to our paper pointed out, incorporating corporate relations even for one year is a very large exercise, and takes many person-years to complete. Hence, it is not viable to incorporate time variations in corporate relations in the patent dataset itself.

[4] The Standard Statistical Establishment List is now called the Business Register.

researchers will be able to analyze using the NBER Patent Data. More specifically, we believe that there are two areas where this concordance may be useful. First, the Census data will allow deeper analysis of smaller and unlisted firms, which are not covered by Compustat. This may be particularly useful to those interested in entrepreneurship, and in questions about patenting behavior in the early stages of their lifecycle. Second, the concordance will help provide a better picture of changes in ownership among patent assignees. This is likely to be important in tracking small patent assignees that begin as a single-unit firm, and eventually get acquired by bigger firms.

This note is intended to provide an overview of how the concordance was developed, and facilitate the use of this concordance by researchers. It is organized as follows. The next section briefly describes the NBER Patent Data and the BR data. This is followed by a description of the methodology adopted to develop the NBER-BR Bridge. An illustration of the use of this Bridge is provided in Section 4. Section 5 presents an analysis of the coverage offered by the Bridge. Section 6 provides a comparison between the NBER-BR and NBER-Compustat Bridges. Section 7 discusses some potential limitations that researchers must be aware of when using the Bridge. Section 8 concludes.

2. Description of Data

2.1. NBER Patent Data

The NBER Patent Dataset is a group of linked data files that contain detailed front-page data on almost 3,000,000 US patents *granted* between January 1963 and December 1999, all citations made to these patents between 1975 and 1999 (over 16 million) and a reasonably broad match of the patents to COMPUSTAT. Some of the variables in the dataset include the patent number, year of patent application, year of patent grant, number of citations, patent assignees (i.e. firms that are assigned the patents), the associated patent classes etc.

Briefly, the NBER Patent Dataset contains five main data files. The central data file (pat63_99) is a *patent-level* dataset that includes front-page information on all the utility patents granted during that period. The citations file, cite75_99, includes all citations made by each patent granted during the period 1975-1999. The coname file provides the full assignee names. The fourth file, the inventor file (ainventor) contains the full names, city and state of each of the inventors listed on each patent. Finally, the NBER Patent Dataset contains a link to Compustat data in the match file that matches the assignee to the Compustat CUSIP identifier. This match is done for about 4,900 firms and is based on the 1989 universe of firms. A very detailed description of this dataset is provided in Hall et al, 2001.

2.2. BR[5]

The US Census Bureau has a mandate, authorized by USC Titles 13 and 26, to provide a current and comprehensive database of US business establishments and companies for statistical program use. It is under this mandate that the Census Bureau develops and maintains the Business Register (BR). Briefly, the BR is the US Census Bureau's master business list. The records in the BR are constantly updated using information from payroll tax forms, and corporate income tax forms. In the case of multi-location firms, these data are augmented with information collected from various Censuses and surveys of the US Census Bureau (Davis et al 2006). The Census Bureau creates snapshot files every year that are available for analysis. Historically, these files go back to 1975.

The primary use of the BR is to identify target populations for business statistics programs within the Census Bureau (e.g. the target list for the Economic Censuses). Its secondary function is to provide data for various statistical products developed by the Census Bureau, and as an important source for economic research at the Census Bureau and the various Census research data centers.

The BR identifies different types of statistical units. The two most relevant for our purposes are the establishment and enterprise. *Establishments,* typically individual locations or

[5] Given the scope of our work under this approval, we had detailed firm data only for manufacturing firms. Hence, we originally envisaged the Bridge to be a concordance based on names from the BR, but limited to manufacturing firms. However, the absence of SIC industry codes in the patent dataset made it impossible to identify (using the patent dataset) if a specific patent assignee was in the manufacturing sector. Hence, instead of the limited Bridge originally envisaged, we constructed a Bridge covering all sectors of the US economy.

plants of a firm, *are the lowest level of aggregation in the BR*. An enterprise (or a firm) is a collection of one or more establishments. Enterprises are classified into single establishment enterprises (those that operate just one establishment) and multi-establishment enterprise (those that operate two or more establishments). Multi-establishment enterprises may have very simple structures with no subsidiaries with the same name for all establishments or could have a more complex structure with one or more subsidiaries, each with a different name. In order to keep track of the ownership structure among multi-establishment enterprises, the Census Bureau groups all establishments belonging to a single multi-establishment enterprise under a firm ID (ALPHA). In addition to information from payroll tax records, the Census Bureau uses an annual Company Organization Survey to track the establishment structure of multi-establishment enterprises (Davis et al 2006). This survey covers all multi-establishment enterprises and a sample of single-unit enterprises.

The three main types of data in the BR that we use for developing the Bridge are the business name, the postal address and the industry classification. It is important to note that the business name and postal address are not available in other Census datasets such as the economic Censuses. Hence, any matching to external databases that involve names and/or postal addresses must rely on the BR.

3. Creating the NBER Patent Data-BR Bridge

The NBER Patent Data-BR Bridge provides a concordance between the patent assignee (identified by the assignee number) and a Census firm (as identified by a constructed variable, MTCHVAR, described below) *for every year that a patent assignee is matched to a Census firm*. Broadly, this concordance was based on the matching names from the two data sources. Before we detail the steps used to construct the Bridge, we provide an overview of our approach, and discuss a few key decisions that users of the Bridge should be aware of.

3.1. Overview of our approach

Matching names within a single data source is often a difficult exercise. Having to do that from two different data sources is even more challenging. Some of the common difficulties in name matching include misspellings, variations (e.g. ABC Inc and ABC Incorporated) and name changes.[6] The same difficulties extend to "blocking variables" such as the city and state (e.g. Watertown and E. Watertown).[7] Hence, trying to obtain exact string matches is not a very fruitful exercise. Not surprisingly, all name-matching algorithms focus on approximate string matching, not exact string matching. The method of approximation varies across different software packages, but generally they attempt to minimize one or more "distance metrics", typically based on the similarities and differences in the letters of the two names. In addition, most name-matching algorithms complement string similarity with phonetic similarity to identify potential matches.

Approximation, however, comes at a price. While it alleviates the impact of problems such as misspellings and increases the probability of finding a correct match, approximation also increases the probability of finding false matches (e.g. "Rob Smith Inc" would very likely be grouped with "Bob Smith Inc" even though they may be different entities). Hence, most algorithms allow the user to choose their own level of approximation based on their individual trade-off between the probability of finding a correct match and the probability of finding a false positive. Furthermore, they also allow the use of additional information such as the postal

[6] Name matching is a field of study in its own right, and in this note, we cannot hope to describe all the challenges associated with name matching.

[7] Blocking variables are variables used to limit the scope of name matching. For instance, if the city is used to block, we look for entities that have same or similar names within the same city.

address to increase the probability of finding a correct match. Nevertheless, even with the most stringent rules and additional information, name matching will still remain an exercise in approximation.

Our approach to matching assignee and firm names was based on a two-stage combination of a standard name-matching algorithm followed by the application of decreasingly restrictive sets of user-defined string matching rules. *Throughout our approach, we placed much greater weight on avoiding a false positive than on increasing the probability of a correct match.*[8] In the first stage, we obtained a preliminary set of matched "clusters" (i.e. groups of assignees and establishments with the same name and satisfying some geographic restrictions) using the SAS name-matching procedure DQMATCH. In addition to being able to handle large datasets, this procedure has been used within the US Census Bureau for applications requiring name matching. In the second stage, we used the Employer Identification Number (EIN) available in the BR along with a series of user-defined string matching rules (discussed below in detail) to obtain the final matches.

3.2. Unit of Matching

Patents are granted to assignees, but data in the Census datasets are typically at the establishment level. In general, there is no one-to-one correspondence between assignees and establishments. It is not necessary that inventing locations for a firm be the same as its "business locations" where it performs manufacturing or delivers its services. This is not a big problem for small firms that tend to have only one location; however, it is an important concern for larger firms. Another problem is that the assignee name on a patent depends on the patenting policies of individual firms. For example, assignees may be individual locations (e.g. the R&D lab) of a multi-location firm, the headquarters of a multi-location firm, or simply a

[8] Clearly, this is not without its limitations. We discuss some of these in section 7.

patent-holding subsidiary that manages all the patents of the firm. Again, this is likely to be a bigger concern for large multi-establishment firms than for single unit firms.

Hence, we build the primary concordance in the Bridge between assignees and firms, and not between assignees and establishments.[9] In addition to addressing the problems above, this decision is also consistent with patenting being a firm-level activity, whose benefits extend beyond the inventing location.[10] *This implies that a firm could be matched to more than one assignee in any given year.*

We use the establishment identifier in the BR to identify firms. Typically, single establishment firms have the same identifier for the firm and the establishment. Multi-establishment firms, in contrast, have a firm identifier common to all establishments, which is then modified by adding subsequent digits to obtain identifiers for the establishment. As our primary linking variable, we construct a variable, MTCHVAR, equal to the establishment identifier in the case of single establishment firms, and equal to the firm identifier in the case of

[9] Note, however, that we do use the name and geographic information of establishments for developing the assignee-firm concordance. Further details are provided later in this section.

[10] Nonetheless, note that our matching does not preclude researchers from pursuing a more detailed establishment level, albeit more customized match for their own projects (e.g. if researchers want to examine the effect of geographic proximity between patenting and manufacturing locations, they could start with our match and use the state codes or city names to assign plants and patenting locations to geographic "cells"). We also created a secondary establishment-level match in cases where we could uniquely identify an establishment name and id with a patent assignee. However, for reasons discussed above, we would advise researchers to exercise caution when using this secondary match.

multi-establishment firms.[11] Changes in firm identifiers typically happen due to corporate control events such as acquisitions or divestitures.

3.3. Timing of the Match

The NBER Patent Dataset offers two potential choices of timing –the application year and the grant year. The application year denotes the year in which the application for the patent was filed. The grant year refers to the year in which the patent office granted the patent after a review of the patent application. The application year is not only closer to the true date of the invention, but is also less affected by the uncertainties of the patent approval process. The grant year, on the other hand, depends on the length of the approval process, which may vary depending on various factors including the complexity of the technology involved, and the quality of the initial patent application.

In constructing the Bridge, however, we allow for the flexibility of using either the application year or the grant year for matching. Specifically, the Bridge attempts to match each assignee to a BR firm in all the years that the firm appears in the BR. For instance, if the firm ABC Inc appears in the BR during 1977-83, this firm would be matched to assignee ABC Inc for each of these seven years (assuming no mismatches).[12] This allows researchers to use either the grant year or the application year for matching patent data with firm data. For instance, if ABC

[11] We did not have access to the Longitudinal Business Database (LBD). Hence, we could not use the firm identifier in that database. It is, however, easy to link to the firm identifier in the LBD because the BR is the primary source for the LBD.

[12] There are some reasons why an assignee might not be matched to a firm in a year even though both the assignee and the firm exist in the two databases in the same year. Typically, this is due to problems with the name. We discuss this in more detail in section 7.

Inc applied for patents only in the years 1977 and 1979, one could use the Bridge to merge patent data during these years with firm data for these years from the BR. Similarly, if the interest was in the grant year, and say, ABC Inc was granted patents in 1977 and 1980, one could limit the patent data-firm data merge to only these years. A detailed illustration using the application year is provided in section 4.

3.4. Geographic Variables used for Matching

The matching algorithm uses geographic information to block the scope of name matching. The NBER Patent Dataset does not contain geographic information (e.g. city and state) of the patent *assignee*; however, it does provide this information for all *inventors* on the patent. Hence, we decided to block the name matching on geographic information of *any* of the patent's inventors. Specifically, we used either the state or the city and the state for blocking the name matching. For example, if we blocked on city and state, the algorithm would look for BR establishments with the same (or similar) name as the assignee that are located in the same city and state as any of the patent's inventors. We could not use the complete postal address because this was not available in the NBER Patent Dataset.

Though it may seem that not having geographic information for the assignee would be a major limitation, it is not the case. As discussed earlier, it is not necessary that the location of the assignee be the same as the inventing location. For instance, all of IBM's patents are assigned to IBM at Armonk, NY even though IBM has many research labs across the country.[13] Hence, between having geographic information on the assignee and on the inventors, the latter is preferable.

We used two levels of blocking. The city and state were used as blocking variables to obtain a more reliable level of matching (reliability levels are discussed below). The second

[13] This example refers to a real company and is based on public data.

level used only the state as the blocking variable.

3.5. Scope of the Bridge

The BR data are not available before 1975, and 1997 was the last Census that these authors had access to. Furthermore, the NBER Patent Dataset does not contain any patent or inventor data after 1999. Hence, the first year of the match was chosen to be 1975, and the last, 1997. The BR covers all industries, and so does the Bridge.[14]

Only US assignees that were not individuals or governments were selected for matching. The primary reason for this restriction was that the BR does not cover individuals or foreign firms without establishments in the US.[15]

[14] A majority of patents, however, belong to manufacturing firms. Of all patents applied between 1975 and 1997 that eventually matched to a BR firm, approximately 70% belonged to manufacturing firms.

[15] While this restriction largely limits the patents to those belonging to firms, it also includes patents of research institutes and universities. Hence, researchers interested only in firm patenting should try and exclude these patents.

3.6. Levels of Reliability

While the DQMATCH procedure was very efficient at combining similar names into clusters, many of these clusters contained multiple names, and were not directly useful for creating the Bridge. Hence, we processed these clusters in a second step by applying certain rules to obtain a "cleaned" name, which was then used for obtaining the final matches. Any assignee and an establishment of a firm whose "cleaned" names matched exactly were treated as a match, with the assignee number being "bridged" to the MTCHVAR of the establishment. A tiered approach to cleaning was used. First, the most restrictive rules were used to define a match. Any assignees that were matched in this step were identified as a match and excluded from subsequent processing. We then applied a set of less restrictive rules to the remaining assignees to obtain additional matches, but with a potentially lower degree of confidence in the match. The resulting Bridge had matches in five "reliability codes" depending on the geographic information used for match (city vs. state) and the rules applied for cleaning the names.[16]

User-defined set of rules	**Geography variable(s) to block in initial DQMATCH runs**	
	City and State	State
Set 1 (most restrictive)	A	D
Set 2 (less restrictive)	B	E
Set 3 (least restrictive)	C	

We now describe the matching process in detail.

[16] In the actual Bridge file though, for reasons of programming choice, these codes are listed differently (A:3;B:4;C:43;D:5;E:53).

3.7. The Matching Process

3.7.1. Step 1: Obtain the names of patent assignees

We began by excluding all *patents* applied for before 1975 or after 1997 (using the Patents File in the NBER Patent Dataset, apat63_99.dta). We then excluded all patents that did not belong to US assignees (i.e. country code was set equal to "US") or to individuals or governments (i.e. assignee code was set equal to 2). Using the patent assignee number, we then merged this list of patents with the names of the patent assignees from the Company Name File (aconame.dta) in the NBER Patent Dataset. We then retained all unique patent assignee names along with the corresponding assignee number.

3.7.2. Step 2: Obtain geographic information on each patent assignee

As described earlier, geographic information on patent assignees was based on the inventors associated with their patents. We began by identifying all patents associated with the set of assignees identified above in Step 1. We then identified all inventors associated with that set of patents. This was achieved by merging the list of patents with the Inventors File (ainventor.dta) using the patent number, and excluding inventors that were not matched to this list of patents. This resulted in an inventor-level file that contained the patent number, the patent assignee number and name, the inventor ID, and the city and state of the inventor. Using this file, we then created two files – (i) by retaining every unique combination of assignee number (or equivalently, assignee name), city and state (asgcity.sas7bdat), and (ii) by retaining in the unique combination of assignee number and state (asgst.sas7bdat). These two files were used to block the name-matching algorithm at differing levels of reliability.

3.7.3. Step 3: Obtain the name and geographic information on each BR <u>establishment</u> for each year from 1975 to 1997

In this step, we created a data file (one for each year) containing the name, city and state

of each BR establishment based on the BR. Note that even though our final objective is to match the patent assignees to BR *firms*, this file contains *establishment*-level information. There were a number of reasons for using the lower level of aggregation. At the outset, it is difficult to identify unambiguously which establishment (of a multi-establishment firm) should be chosen as "the firm". The BR does not separately identify the headquarters of firms from other locations. More importantly, not all establishments of the firm may have the same name as the firm or be located in the same place as the headquarters. Hence, it is entirely possible that a firm with the name ABC Inc based in Cupertino, CA has patents that are assigned to its subsidiary named XYZ Inc in Rochester, NY. Using firm information in this case would not result in a match; it would be correctly matched if establishment level information were used.

In this step, we created one file for each year, which is in contrast to Step 2 where only a single data file containing assignee information for all years was generated. There were three reasons for this asymmetry. First, as outlined earlier, matching a patent assignee to a firm in all the years that the firm appears in the BR allows researchers the flexibility of using their choice of year to match patent data with firm data. Second, in addition to providing a "contemporaneous match" (e.g. where a patent assignee is matched to a BR firm in the year of a patent application), this also allows researchers to develop an "ever-match" (where a patent assignee is matched to a BR firm sometime during the 1975-97 period irrespective of when they applied for a patent). The final reason was logistics. Over the 23-year period, the BR contains well over 300 million observations. Even though the number of patent assignees is relatively small (approximately 76,000), the processing time required for developing the initial set of matches increases more than linearly in the total number of observations. Doing the matching year by year considerably reduced the number of observations to be analyzed at any one time (from hundreds of millions to less than a tenth of that), thus expediting the execution of the

name-matching algorithm. In the implementation of the code, the size of the datasets meant that we had to split even the annual files into a number of smaller ones so that the system memory and processing constraints would not bind. We did this by one-digit industry, and where even that was large (e.g. retail), we used two-digit industry.

3.7.4. Step 3: Construct the Census firm identifier variable, MTCHVAR, for each BR establishment

The identifier is defined in Section 3.2.

3.7.5. Step 4: Append the patent assignee and establishment information

In this step, we created the files that form the basis for the name-matching algorithms. We appended the patent assignee information from step 2 with the establishment information from step 3. Broadly then, there were two kinds of files – one containing both the city and state (*city.sas7bdat) and the second containing only the state information (*st.sas7bdat) – with the following structure: 2-3 columns containing the name, city and state or just the name and state with the initial 76,000 or so rows containing the patent assignee information followed by BR establishment information in the remaining rows.

3.7.6. Step 5: Generate preliminary clusters of matches

In this step, we applied the DQMATCH procedure in SAS to the files generated in step 4 (e.g. *city.sas7bdat). We generated two sets of clusters - (i) based on the city and state as the blocking variables and (ii) based on the state alone as the blocking variable.

Because we chose to place a greater weight on avoiding false positives, we set the "sensitivity" in the DQMATCH procedure to the highest permissible level. While we do not know the inner workings of this procedure, we understand that setting a higher level of sensitivity increases the threshold level of similarity that must be met before two (or more) names are called a match.

Note that this procedure is different from the usual merger of data from two different

data sources based on one or more linking variables. Here, we have only one data file that contains the name and other information from both data sources. The objective of this procedure then, as applied here, is to identify *duplicates* within the single data file. Hence, a cluster generated by this program could contain only establishments from the BR (and no assignees from the patent dataset) or only assignees (and no establishments from the BR) or a combination of many assignees and many establishments. We dropped all clusters with no assignees in them, and those with more than one assignee in them.

At the end of this step, we had two resulting sets of clusters for each year from 1975 to 1997. The first set contained clusters of exactly one patent assignee and one or more BR establishments in the same city and state that had the same or similar names. The second set contained clusters of exactly one patent assignee and one or more BR establishments in the same state that had the same or similar names.

3.7.7. Step 5: Apply the first set of user-defined string similarity rules to clusters from step 4

Even with DQMATCH set at the highest level of sensitivity and the use of city and state as blocking variables to match names, the resulting clusters usually contained too many establishments to be of use without further modifications. In line with the objective of reducing the probability of false positives, we applied fairly strict string similarity rules to the clusters obtained from step 4 to identify matches between assignees and establishments. As explained above, the names of the assignees and firms in these clusters were "cleaned" by applying rules depending on the reliability code. Any assignee and an establishment of a firm whose "cleaned names" matched ***exactly*** were treated as a match, with the assignee number being matched to the MTCHVAR of the establishment.

We began with the strictest set of similarity rules, and then sequentially applied a series of decreasingly strict string similarity rules.

Name cleaning rules: Set 1 (Most restrictive)

(a) Suffixes INC/INC./INCORPORATED/INCORPORATED./INCORP/INCORP/CO INC. were assumed to be identical

(b) Suffixes CORP/CORP./CORPORATION/CORPORATION. were assumed to be identical

(c) Suffixes COMPANY/COMPANY./CO./CO were assumed to be identical

(d) Suffixes LIMITED./LIMITED/LTD./LTD were assumed to be identical

(e) The word ASSOCIATION was assumed to be the same as the word ASSN

(f) The words MANUFACTURING and INTERNATIONAL were assumed to be the same as the words MFG and INTL respectively

(g) If the first word in a name was THE, that word was dropped

(h) The symbols ",", "%", "(" and ")" were all dropped

(i) The symbols "&" and "+" were treated as the word AND.

At the end of the step, assignees could be in one of two bins – those that get matched to a BR firm, and those that do not. Assignees in the former bin go to Step 6 while assignees in the latter bin are processed further in step 7. Those matched could be matched to a unique establishment and hence, a unique firm (for an example, refer cluster 1 in Appendix I), or to a unique firm but more than one establishment (cluster 2, Appendix I) or more than one unique firm (cluster 3, Appendix I).

3.7.8. Step 6: Within clusters that contain assignees matched to a BR firm after step 5, identify establishments with the same EIN as the matched establishment

Besides the establishment identifiers, the BR also provides an EIN used for payroll tax purposes. In order to capture any establishments that may have been wrongly classified as single unit firms (e.g. they have different firm identifiers but the same EIN), we identified all

establishments within the cluster that have the same EIN as the matched establishment(s), and linked them to the assignee. For an example, refer cluster 4 in Appendix I.

3.7.9. Step 7: Apply the second set of user-defined string similarity rules to clusters from step 4, followed by the same procedure as in Step 6

To the assignees not matched in Step 5, we applied the following set of rules.

Name cleaning rules (Set 2): In addition to the rules in Set 1, suffixes INC, CORP, CO, LTD, LLC and PLC, and the symbol "-" between two parts of the name were dropped. Hence, within this code, for instance, ABC INC in Sacramento, CA would be considered the same as ABC CORP in Sacramento, CA. For another example, refer cluster 5, Appendix I. As with Step 5, assignees may or may not get matched to unique firms/establishments. This was then followed by identifying and linking all establishments within the cluster that have the same EIN as the matched establishment(s).

3.7.10. Step 8: Apply the third set of user-defined string similarity rules to clusters from step 7, followed by the same procedure as in Step 6

To the assignees not matched in Step 7, we applied the following set of rules.

Name cleaning rules (Set 3): In addition to the rules in *Set 1*, the following rules were applied:

(a) Spaces between names were dropped as were the symbols "-" and "/" between two parts of the name.

(b) The letter "S" was dropped from the names (to identify names that differ only by the letter "S", mostly in the case of plurals)

(c) The word AND was dropped

(d) Drop the last words of a name if they were either OF or FSC.

For an example, refer cluster 6, Appendix I. As with Step 5, assignees may or may not get matched to unique firms/establishments. This was then followed by identifying and linking all

establishments within the same cluster that have the same EIN as the matched establishment(s).

All assignees that remained were classified as non-matches and excluded (e.g. cluster 7, Appendix I).

4. Using the NBER-BR Bridge –An Illustration

The Bridge provides a concordance between an assignee and a BR firm for every year the assignee matches to a BR firm. It does not contain any firm or patent data. Linking firm data (e.g. employment) and patent data (e.g. patent stock) requires a number of additional steps. There is no single set of steps that can be used universally; the exact steps required depend on the objective of the study in question and the Census datasets being used to obtain firm data. Here, we provide an *illustration* of the steps required based on a hypothetical research question – (a) what is the elasticity of productivity to patent stock? We will use application year as the basis for aggregating patent data, and the quinquennial Census of Manufacturing as the source of firm data. These research questions require the following data for *each firm for each Census year*:

- What is the patent stock (assume 15% annual depreciation) for any firm in a given year?
- What is the productivity of a firm in a given year?

Below, we explain how the Bridge can be used to obtain these data.

Step 1: Obtain a balanced assignee-application year panel and aggregate individual patent data to the patent assignee level

The patent dataset provides data on individual patents, which must then be aggregated to the assignee level before the Bridge can be used. Though aggregation is a straightforward exercise, the complication is that not all assignees apply for patents every year, resulting in an

unbalanced assignee-application year panel.[17] Hence, we balance the panel by filling in the years (tsfill command in Stata) in which the assignees do not apply for a patent. We then aggregate by counting the number of patents applied by an assignee till a given year, which gives the total undepreciated patent stock. The total depreciated patent stock is obtained by appropriately depreciating the prior year's patent stock and adding the current year's applications.

Step 2: Aggregate establishment data to the firm level

We aggregate establishment level input and output information to the firm level, by adding up inputs and outputs of individual establishments within the firm. Firm level productivity then can be measured using any of a number of methodologies proposed in the literature (eg Solow residual, a simple OLS-Fixed Effects specification, or Blundell and Bond 2000). This results in a firm-year level dataset with firm output, inputs and productivity for every Census year that the firm is covered by the Census.

Step 3: Merge the balanced assignee-application year data from step 1 with the Bridge

Step 4: Aggregate assignee data to the firm level

At the end of step 3, we have a data file that contains the patent stock by assignee along with the corresponding firm identifier (MTCHVAR) for every year the assignee is matched to a BR firm. Because we chose the application year as the basis for aggregating patent data, the patent stock will reflect all applications made till that year. *A firm may be matched to multiple assignees, and hence, we add the stocks of all assignees that are matched to a MTCHVAR in a given year.* This

[17] This is problematic because the Census of manufacturing has data only every five years. Hence, if a firm did not apply for a patent in a Census year, it will not be observed in the merged data.

gives the total patent stock for each firm for each year that the firm is matched to at least one assignee. We then collapse the data to the firm-year level, the same level of aggregation as that from step 2.

Step 5: Merge the data file from step 4 with that from step 2 to get the final matched patent data-firm data

This data file can now be used to examine the research question.

For further illustrations, refer Appendixes II and III. Appendix II provides three different examples of how the firm data may be linked to assignee data. Appendix III illustrates how acquisitions may be treated in the merging process.

5. Analysis of the Bridge Coverage

We analyze coverage using two types of matches. The first, "ever-matched", refers to a patent assignee-BR firm match irrespective of the year of patenting or the year of occurrence in the BR. Hence, if an assignee applied for a patent only in 1981 (i.e. did not apply for patents at any other time), and was matched to an BR firm in the years 1984 and 1985, then, this would be classified as an "ever-match" even though the year of the application and the firm's appearance in the BR do not coincide. The advantage of this match is that timing errors such as those caused when a firm patents as a non-employer firm (and hence, is not included in the BR), and eventually has employees (and hence, appears in the BR) are addressed. However, the disadvantage is that it increases the potential for mismatches, and in addition, does not allow tracking of ownership changes.

The second type of match is a "contemporaneous match", which captures inter-temporal variations in the links. There are two potential candidates for such a match-one based on the grant year and one based on the application year. We chose application year as the basis to define a "contemporaneous match"; so, it would be defined as a patent assignee-BR firm match in the application year of patenting. Hence, in any given year, a contemporaneous match exists

only if the assignee applied for a patent during that year, and can be found in the BR. When matched, this is clearly more accurate than the "ever-match", but the chances of a non-match occurring due to timing errors such as the one discussed above are higher.

Table 1 presents the coverage statistics for various populations. Of the 2.92 million patents in the NBER Patent Dataset belonging to 175,115 assignees, about 40% belonging to about 57,600 assignees were matched at least once ("ever-match") to a firm in the BR. Within the relevant population of patents applied between 1975 and 1997 by US assignees that were not individuals, universities or government agencies, 90% of patents, and 63.7% of assignees were "ever-matched". Using the contemporaneous match criterion, about 80.6% of patents were matched to a BR-firm in the year of application. 64% of assignees were matched to a BR-firm in the year of application.

Figure 1 plots the year-by-year coverage for the contemporaneous match. Patent coverage hovers around 80% in most years with the exception of 1988 when it is 65%. This is because a large part of the BR firm data for this year was not available. When measured using citation-weighted patents, the annual coverage is similar, varying from 78% to 86% (with the exception of 1988). Assignee-level coverage varies from about 53% in 1988 to about 70% in 1977, with an average of about 64%. Hence, it appears that the contemporaneous match is reasonably uniform over time. The "ever-match" coverage (not presented here), expectedly, shows a mild declining trend over time from about 80% of assignees matched in the earlier years to about 70% in 1997.

Table 2 presents the coverage by technological category. The NBER Patent Dataset provides six technological categories (based on over 400 US Patent Classes). While there is some variation across technologies (about 75% of patents contemporaneously matched in Drugs and Medical to 84.5% in Electrical and Electronics), no technological category appears to be

unreasonably under-matched or over-matched.

In order to identify any systematic biases in the matching process, Table 3 compares the matched and non-matched assignees in the NBER Patent Dataset. Not unexpectedly, the matched assignees are larger (about 1.73 patents per year compared to 1.27 patents per year for the non-matches). They also appear to patent for longer periods of time as evidenced by the bigger difference in the "ever-matched" criterion - 14.42 patents over 1975 to 1997 for the matches compared to 2.81 for the non-matches.[18] Finally, there appears to be almost no difference in the average quality of the patents between the matched and non-matched assignees. The average number of forward citations per patent is about 6.2 for both types of assignees. Similarly, the number of claims per patent is only slightly different (13.60 vs. 13.94).

Table 4 examines differences across reliability codes. A majority of patents (81%) and assignee-years (61%) are matched under the highest reliability code, A. The next biggest sets of matches are under reliability code D (10% of patents and 27% of assignee-years). Though these reliability codes vary in the level of geographic blocking (city and state for A, and state for D), both codes use the strictest set of rules for cleaning and matching the names. Matches under other reliability codes form a small fraction of all matches. As expected, larger assignees tend to be matched more reliably than smaller assignees. This can be seen in the decreasing number of

[18] While part of this difference is likely due to genuine reasons such as the smaller firms not being included in the BR, some of this difference could be due to inaccuracies in the matching process itself – e.g. a larger firm may have more establishments, and hence have a greater probability of being matched. Nevertheless, given that almost 80% of patents, and 64% of assignees were contemporaneously matched, the errors due to non-matches are likely to be relatively small.

patents per year (or citation-weighted patents per year) from reliability code A to reliability code E.

Table 5 examines differences in the coverage across assignees of different sizes. We are able to present coverage statistics only for the smallest assignees (defined either as assignees that have less than 20 patents or less than 50 citation-weighted patents). Not surprisingly, coverage for the smaller assignees was below the average for all assignees. The difference is much larger in terms of patents but not as high when assignee-years are used for comparison. Coverage statistics for larger assignees were so high that they posed significant disclosure risk. Hence, they are not presented.

6. Comparison with the NBER Compustat Bridge

6.1. Strengths of the NBER-BR Bridge

The primary purpose of the NBER-BR Bridge is to provide a link between patent data in the NBER Patent Dataset and firm data available in Census datasets. Because the US Census Bureau invests a considerable amount of time and effort in developing and maintaining its BR, the Bridge provides two advantages over the existing NBER-Compustat Bridge - greater coverage and more accurate tracking of ownership changes.

The NBER-BR Bridge provides only a modest improvement in terms of patent coverage (the NBER-Compustat link covers 50-70% of US patents and 30-50% of all patents vs. about 80% in the NBER-BR Bridge for US patents).[19] However, it offers a considerable improvement in terms

[19] Note that we combine data from the Compustat-BR Bridge available at the Census with our NBER Patent Data-BR link. Thus our link incorporates listings after 1989 as well as changes in patent stocks though acquisitions and sales for listed firms after 1989. Accordingly, our coverage of patenting by listed firms is more extensive than would be available with the

of the assignees covered (4,900 vs. 57,000). This is logical since Compustat covers only the larger firms in the economy while the BR covers all firms that pay payroll tax. The use of Census data also provides a much more accurate picture of patenting. Table 6 compares the role of patenting firms in the economic manufacturing sector as measured using the NBER-BR Bridge with that computed using the NBER-Compustat Bridge. The share of patenting firms in net sales measured by Compustat is similar to that of the share of value added obtained from the Census data. However, as expected, Compustat overstates the proportion of patentees in every SIC2 industry. On average, a little over 25% of all Compustat firms are patentees but less than 6% of all Census firms patent.

The second and more important benefit is a spillover of the Census Bureau's investments to track ownership structure among establishments in the BR. The NBER-BR Bridge offers a more accurate tracking of ownership of assignees than the NBER-Compustat link. This is likely to be particularly important since patent owning firms appear more likely to change ownership than other firms. Indicative evidence of this comes from Table 7. This table compares single-unit establishments that are matched to a patent assignee in 1977 with those establishments that are not matched to a patent assignee in the same year. It tracks these cohorts over the next four Census es, and presents the fraction of establishments that exhibit a change in ownership. It is evident that the cohort of establishments that are matched to a patent assignee exhibit a much greater propensity to change ownership than those that are not matched to a patent assignee.

---

existing Compustat-patent data link, which (as discussed above) linked based on ownership links, firm and assignee names for 1989.

6.2. Strengths of the NBER-Compustat Bridge

The currently available NBER-Compsutat Bridge has some important advantages over our NBER-BR Bridge. First, it allows researchers to exploit the more widely available Compsustat dataset for research work; access to US Census data as well as our concordance is more restricted (as noted in 7.1 below). One advantage of this is that the Compustat link is subject to refinement by researchers working on this data. In this connection, though access is more difficult, we expect our bridge file to be similarly refined and improved by other researchers at the Census Bureau.[20] Second, the link can be used to exploit external data such as executive compensation data, and other data collated on listed firms. Finally, Compustat has information on foreign subsidiaries, which is unavailable in the Census databases. Thus, potential impact on foreign operations of patenting can be studied using Compustat data but not with Census data alone.

Researchers who have access to Compustat data and obtain access to the Census data, can combine information from the Compustat data and the Census data, using the Compustat-BR link available at the Census (for 1979 to 2005), along with our NBER -BR Bridge.

7. Limitations of the Bridge

As with all concordances, the NBER-BR Bridge has a number of limitations. We discuss these below. We do not claim that these are the only limitations; they are the ones that we could identify during our use of the Bridge. We expect that future work will not only identify more limitations but also address some of them.

[20] We are aware of two other projects that have requested access to our data. We expect this number to increase in the future. The authors hope to themselves work on Census projects in the future utilizing this link.

7.1. Matching and Patent Ownership

It is not necessary that a firm matched to a patent assignee own all the patents granted to that assignee. The patent documents, as well as the NBER Patent Data, only provide the name of the original assignee. Hence, if an assignee reassigns or sells any of its patents to another firm, the NBER Patent Data will not capture this event. This limitation *partially* carries over to the NBER-BR Bridge. If, along with all the patents, a firm sells the establishment that was assigned the patents, and that establishment retained its name, then the Bridge will reflect the ownership change. The Bridge will also be accurate if the whole firm changes ownership. However, if a firm reassigns or sells only its patents without selling any associated establishment(s), then the Bridge will be inaccurate. Unfortunately, this is a limitation of most studies in this literature, and to our knowledge, there is no comprehensive database of patent reassignments that may help address this limitation.

7.2. Cyclicality in the quality of ownership linkages

As noted by Jarmin and Miranda (2002), the quality of ownership linkages, and hence, the tracking of multi-unit/single-unit firms is considerably better in the Economic Census years than during the inter-Census years. This is because information on ownership is collected from all establishments only in the census years, but only from a subsample of firms in the inter-Census years.

7.3. Non-employers

The BR does not cover firms that do not have any employees (e.g. an entrepreneur in a proprietorship). Hence, any assignees that are not incorporated and have no employees will not be matched in the Bridge.

7.4. Foreign subsidiaries of US firms

The US Census datasets track information only for US establishments. Thus, information on foreign subsidiaries of US firms is unavailable. This implies that the impact of having foreign operations cannot be studied using our Bridge and Census data alone.

7.5. US subsidiaries of foreign firms

Our Bridge includes US subsidiaries of foreign firms. Data to identify subsidiaries of foreign firms in the Census are currently unavailable to external researchers, though this information is tracked through Census questionnaires. Therefore, if the researcher wishes to exclude foreign subsidiaries, they would have to identify these assignees using their own data sources.

7.6. Non-matched assignees

As explained above, about 37% of the relevant population of assignees is not matched to a firm in the BR. While we do not know all the reasons for assignees not being matched, some of the possible reasons include the following.

- Some assignees may not have any employees, and hence, do not appear in the BR. This could happen for one-person start-ups that appear as a corporate assignee in the patent data, but have no paid employees, and hence, do not appear in the BR. This could potentially also arise if a large company incorporated a zero-employee subsidiary as a holding company for its patents.[21] The Bureau has recently assembled a new Integrated Longitudinal Business Database that includes information from a register of non-employer firms. In future work, researchers could extend our link to cover the non-employer register as well, which could improve coverage of assignees significantly.

[21] We understand from Census personnel that coverage of zero-establishments is not uniform across firms.

Such a link could also examine the transition of small non-employer or self employed inventors to corporate status.

- The name of an assignee may be different from that of the firm in the BR. It is possible that the names are completely different. However, it is much more likely that the string matching rules are too strict to overcome data entry errors in one or both of the names (e.g. cluster 7, appendix I).
- Some assignees may have similar names, and hence, classified into the same clusters by the DQMATCH procedure. Since we dropped all clusters that have multiple assignees, these assignees would be excluded.

7.7. Broken matches for an assignee

Some assignees may not be continuously linked to a firm in the BR even though the firm exists in the BR. There are two potential types of broken matches. The first, and observable, is an assignee that matches to a firm in some years but does not match to a firm in some of the intermittent years. The most likely reason for this is data entry error in the names that results in different names in the two datasets.

The second type, and unobservable, is an assignee that matches to a firm for a few years but does not match to a firm after that. This may happen when assignees change their names for patenting purposes. These events are not captured by the patent dataset, and hence will result in inaccuracies in the Bridge. More specifically, suppose assignee John Smith Inc was matched to a firm with the same name in the BR. If in 1994, it changed its name to J Smith Inc for patenting purposes but did not do so in the BR, the assignee will not be matched to any firm in the BR in and after 1994.

7.8. Assignee linked to multiple firms

Some assignees may be linked to more than one firm (e.g. cluster 3, Appendix I). In such cases, the patent stock of the assignee will be allocated to all the firms.

7.9. Changes in the names of assignees after acquisitions

It is possible that the names of assignees are changed after they are acquired by other firms. This may cause inaccuracies depending on whether the acquired assignee continues to remain an establishment of the acquired company and on whether the acquired assignee continues to patent after the acquisition. Scenarios 3 through 5 in Appendix III illustrate the potential errors under three possible scenarios.

7.10. Accessibility

Census data as well as the NBER-BR bridge can be accessed only at Census Research Data Centers, which requires that researchers go through a proposal submission and review process.

8. Conclusion

The efforts of Hall, Jaffe and Trajtenberg to construct the NBER Patent Dataset and link patent data to firm data have significantly improved our understanding of innovation in general, and patenting in particular. The result of this project, the NBER-BR Bridge, extends those efforts by enabling researchers to link patent data with firm data at the US Census Bureau. The Bridge also benefits from the considerable investments made by the US Census Bureau in tracking the ownership structure of establishments in the US "employer universe". These spillovers from the original efforts of Hall, Jaffe and Trajtenberg (2001), and from the investments made by the Bureau have enabled us to extend the coverage of firm data-patent data links beyond those available in the NBER Patent Dataset. It has also enabled us to provide a more accurate tracking of ownership changes among patent assignees.

Notwithstanding these improvements, we recognize that the NBER-BR Bridge will only benefit from the contributions of others who to choose to use it. We invite researchers to build on our efforts.

9. References

Blundell R. and S. Bond.,2000,"GMM estimation with persistent panel data: an application to production functions" *Econometric Reviews*, 19:321-340

Davis, S.J., Haltiwanger, J., Jarmin, R.S., Krizan, C.J., Miranda, J., Nucci, A. & Sandusky, K. 2006. Measuring the Dynamics of Young and Small Businesses: Integrating the Employer and Non-employer Universes, Working Paper CES 06-04, Center for Economic Studies, US Census Bureau

Hall, B.H, Jaffe, A.B. & Trajtenberg, M. 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools.

Jarmin, R. S., and Miranda, J. 2002. The Longitudinal Business Database. Working Papers 02-17, Center for Economic Studies, U.S. Census Bureau.

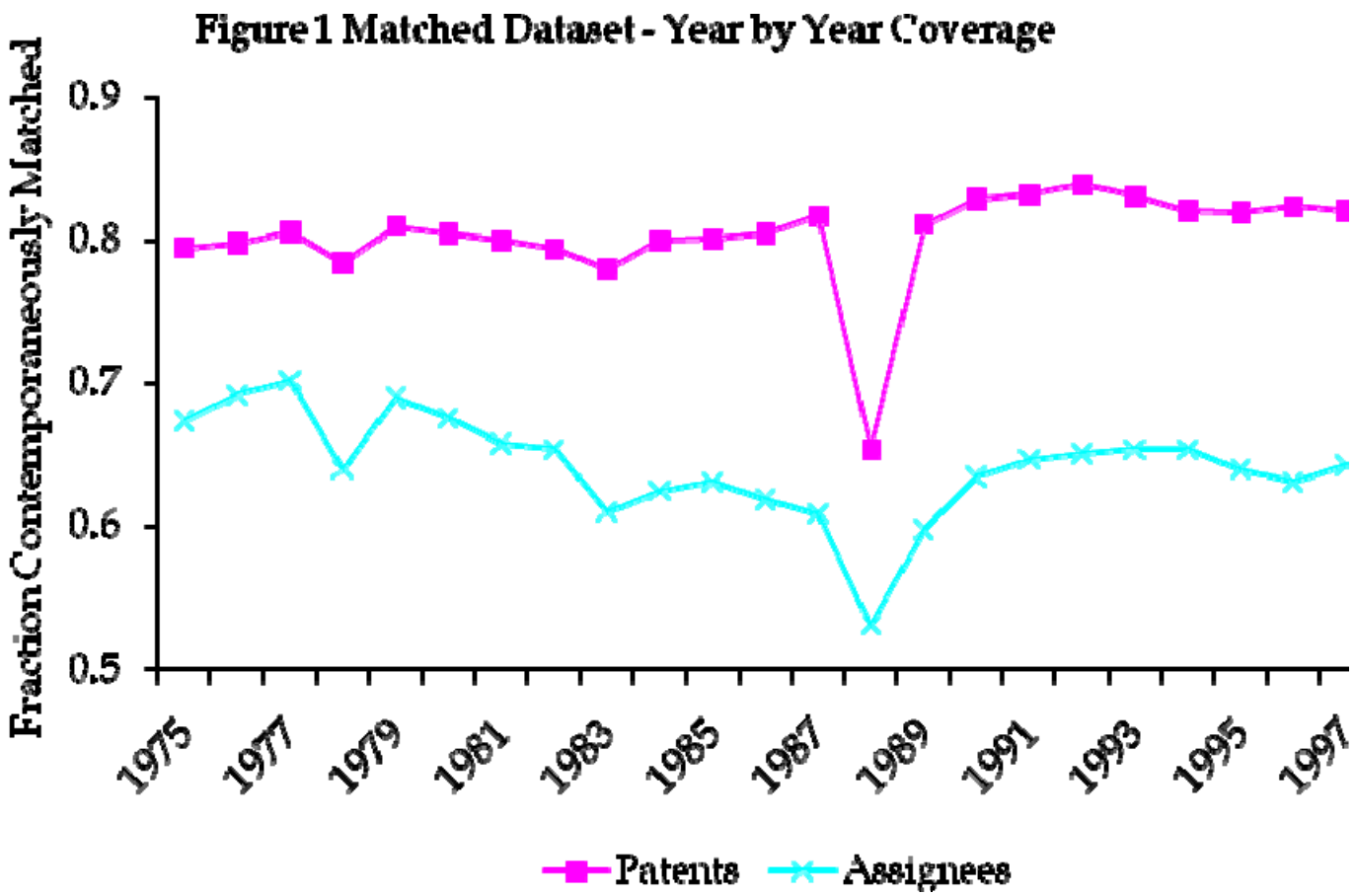
Figure 1 Matched Dataset - Year by Year Coverage
Fraction Contemporaneously Matched
0.9
0.8
0.7
0.6
0.5
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
Patents
Assignees

TABLE 1: MATCHED DATASET – OVERALL COVERAGE

EVER-MATCHED	Patents	Citation-Wtd Patents	Assignees
1963-1999: All Assignees	0.403 (2.92 m)	0.479 (14.0 m)	0.329 (175,115)
1963-1999: US Non-Individual, Non-Government Assignees Excluding Universities	0.899 (1.07 m)	0.908 (6.20 m)	0.599 (95,157)
1975-1997: US Non-Individual, Non-Government Assignees Excluding Universities	0.900 (0.78 m)	0.910 (4.54 m)	0.637 (76,380)

CONTEMPORANEOUS MATCH	Patents	Citation-Wtd Patents	Assignee-years
1975-1997: US Non-Individual, Non-Government Assignees Excluding Universities	0.806 (0.78m)	0.797 (4.54 m)	0.640 (189,835)

Notes: (i) Each cell presents two numbers. The first is the fraction of patents or of assignees covered in the match. The second, in parentheses, provides the total number of patents or assignees in the relevant population. (ii) "Ever-Matched" is defined as a Patent Assignee-BR Firm match, regardless of the year of patenting. A "Contemporaneous Match" is defined as a Patent Assignee-BR Firm *in the **application** year of patenting*. Hence, in any given year, a contemporaneous match exists only if the assignee applied for a patent during that year, and can be found in the BR. (iii) "m" represents million.

TABLE 2: MATCHED DATASET –COVERAGE BY TECHNOLOGICAL CATEGORY

	Ever-Matched			Contemporaneous Match		
	Patents	**Citation-Wtd Patents**	**Assignees**	**Patents**	**Citation-Wtd Patents**	**Assignees**
Chemicals (1)	0.919	0.929	0.699	0.831	0.828	0.688
Computers & Communication (2)	0.925	0.923	0.700	0.833	0.791	0.676
Drugs & Medical (3)	0.856	0.871	0.635	0.750	0.726	0.615
Electrical & Electronic (4)	0.926	0.931	0.719	0.845	0.837	0.714
Mechanical (5)	0.894	0.906	0.702	0.799	0.799	0.694
Others (6)	0.862	0.879	0.674	0.756	0.763	0.662

Notes: The sample covers only patents applied during 1975-1997, and belonging to US non-individual, non-government assignees that are not universities

TABLE 3: COMPARISON OF MATCHED AND NON-MATCHED ASSIGNEES

	Ever-Matched		Contemporaneous Match	
	Matched	**Non-Matched**	**Matched**	**Non-Matched**
Number of Patents	14.42	2.81	1.73	1.27
Number of Citation-Wtd Patents	84.89	14.69	11.71	8.02
Number of Citations per Patent	5.21	5.10	6.22	6.20
Number of Claims per Patent	13.45	13.50	13.60	13.94

Notes: (i) The relevant population is US non-individual non-government assignees, excluding universities. (ii) The match coverage includes patent assignees not in the manufacturing sector. (iii) For "Ever-Matched" analyses, data across all years are used. For "Contemporaneous Match" analyses, data only in the year of application is used. Hence, the number of patents and citation-weighted patents is higher in the "Ever-matched" analyses.

TABLE 4: DESCRIPTIVE STATISTICS BY RELIABILITY CODE
(Contemporaneous Match Only)

Reliability Code	Number of Patents	Number of Assignee-years	Number of patents per assignee-year	Average number of citations to patents (per year)
A	510,036 (81%)	74,460 (61%)	6.85	39.55
B	40,116 (6%)	8,114 (7%)	4.94	29.94
C	8,826 (1%)	3,383 (3%)	2.61	13.43
D	64,366 (10%)	32,503 (27%)	1.98	10.78
E	5,755 (1%)	2,944 (2%)	1.95	11.14

TABLE 5: MATCHED DATASET –COVERAGE FOR SMALL ASSIGNEES

	Small Assignees (1)	Small Assignees (2)	All Assignees
Patents	72.7%	69.8%	90.0%
Assignees	62.5%	61.5%	63.7%

Notes: This table presents the coverage for small assignees. The first row is the fraction of patents ever-matched; the second row is the fraction of assignees ever-matched. Small assignees are defined in two different ways. Small Assignee (1) is defined as an assignee that has less than 20 patents. Small Assignee (2) is defined as an assignee that has less than 50 citation-weighted patents. We are not able to present similar statistics for larger assignees due to disclosure restrictions.

TABLE 6: ECONOMIC ACTIVITY IN MANUFACTURING - SHARE OF PATENTING FIRMS (EXISITING COMPUSTAT VS. REVISED COMPUSTAT VS. CENSUS OF MANUFACTURING)

		COMPUSTAT Existing		CMF	
		Number of firms	**Net Sales**	**Number of firms**	**Value Added**
20	Food and kindred products	20.4%	57.2%	1.9%	60.5%
22	Textile mill products	29.0%	65.9%	3.7%	51.5%
23	Apparel and other textile products	19.8%	71.9%	0.9%	26.5%
24	Lumber and wood products	20.2%	56.6%	0.5%	29.4%
25	Furniture and fixtures	35.2%	59.8%	2.7%	42.6%
26	Paper and allied products	39.9%	69.7%	7.1%	77.2%
27	Printing and publishing	16.7%	49.8%	0.5%	33.4%
28	Chemicals and allied products	21.1%	83.6%	9.2%	86.1%
29	Petroleum and coal products	34.0%	55.4%	7.0%	72.4%
30	Rubber and miscellaneous plastics	30.9%	80.0%	7.6%	59.8%
31	Leather and leather products	24.4%	55.3%	3.4%	40.0%
32	Stone, clay, glass, and concrete products	29.3%	77.4%	2.2%	53.8%
33	Primary metal industries	34.5%	62.5%	6.6%	74.2%
34	Fabricated metal products	33.7%	63.4%	5.2%	48.7%
35	Industrial machinery and equipment	29.4%	75.9%	6.8%	72.5%
36	Electrical and electronic equipment	23.5%	74.8%	12.0%	77.4%
37	Transportation equipment	37.0%	78.1%	5.8%	90.3%
38	Instruments and related products	21.2%	76.3%	17.4%	86.8%
39	Miscellaneous manufacturing industries	19.8%	69.2%	3.8%	43.2%
	ALL INDUSTRIES	**25.7%**	**70.2%**	**5.5%**	**59.3%**

Notes: (i) The numbers represent averages over the period 1977 to 1997. (ii) Due to disclosure restrictions, the statistics for SIC 21 (Tobacco manufactures) have been merged with SIC 20 (Food and kindred products).

TABLE 7: OWNERSHIP CHANGE AMONG SINGLE-UNIT FIRMS

Survived at least till Year	Patenting Status in 1977	
	0	***1***
1982	3.44%	15.04%
1987	8.42%	24.78%
1992	10.69%	29.50%
1997	12.87%	33.28%

Notes: This table examines ownership change among firms that were single establishment firms in 1977. Patenting status is 0 if an establishment is not matched to the patent assignee in 1977 and 1 if it is. Hence, of the single unit firms in 1977 that did not own a patent and survived at least till 1982, 3.44% changed ownership.

APPENDIX I: PROCESSING OF CLUSTERS- SOME EXAMPLES

Cluster 1 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John Smith Furniture Inc, Tempe, AZ	500000	7239	Set 1	Exact Match
(E) Jon Smith Furniture Inc, Tempe, AZ	0714568921	5124	Not Considered	Assignee excluded for Sets 2 and 3 after match is found under Set 1
(E) The John Smith Furniture Company, Tempe, AZ	500000	7240	Not Considered	
(E) John & Smith Furniture Inc, Tempe, AZ	500000	7240	Not Considered	
(E) John Smith Furnitures Inc, Tempe, AZ	500000	7239	Not Considered	

Final Match (Highlighted above): In 1993, assignee 25 matched to BR Firm 500000 with Reliability Code A. Also, assignee matched to a unique *establishment* (establishment ID not presented).

Cluster 2 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John Smith Furniture Inc, Tempe, AZ	500000	7239	Set 1	Exact Match
(E) John Smith Furniture Inc, Tempe, AZ	500000	7240	Set 1	Exact Match
(E) Jon Smith Furniture Inc, Tempe, AZ	0714568921	5124	Not Considered	Assignee excluded for Sets 2 and 3 after match is found under Set 1
(E) The John Smith Furniture Company, Tempe, AZ	500000	7240	Not Considered	
(E) John & Smith Furniture Inc, Tempe, AZ	500000	7240	Not Considered	

Final Match: In 1993, assignee 25 matched to BR Firm 500000 with Reliability Code A. Assignee <u>not</u> matched to a unique *establishment*.

Establishment ID not presented.

Cluster 3 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John Smith Furniture Inc, Tempe, AZ	500000	7239	Set 1	Exact Match
(E) John Smith Furniture Inc, Tempe, AZ	08353218753	8153	Set 1	Exact Match
(E) Jon Smith Furniture Inc, Tempe, AZ	0714568921	5124	Not Considered	Assignee excluded for Sets 2 and 3 after match is found under Set 1
(E) The John Smith Furniture Company, Tempe, AZ	500000	7240	Not Considered	
(E) John & Smith Furniture Inc, Tempe, AZ	500000	7240	Not Considered	
(E) John Smith Furnitures Inc, Tempe, AZ	500000	7239	Not Considered	

Final Match: In 1993, assignee 25 matched to two BR Firms (500000 and 08353218753) with Reliability Code A. Assignee not matched to a unique establishment.

Cluster 4 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John Smith Furniture Inc, Tempe, AZ	500000	7239	Set 1	Exact Match
(E) John Smiht Furniture Inc, Tempe, AZ	0534219781	7239	Set 1	Same Employer Identifier as Exact Match above
(E) Jon Smith Furniture Inc, Tempe, AZ	0714568921	5124	Not Considered	Assignee excluded for Sets 2 and 3 after match is found under Set 1
(E) John Smith Furnitures Inc, Tempe, AZ	500000	7239	Not Considered	

Final Match: In 1993, assignee 25 matched to two BR Firms (500000 and 0534219781) with Reliability Code A. Assignee not matched to a unique establishment.

Cluster 5 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) The John Smith Furniture Company, Tempe, AZ	500000	7240	Set 2	- "The" removed under Set 1. - Company=Inc under Set 2
(E) John & Smith Furniture Inc, Tempe, AZ	500000	5342	Not Considered	Assignee excluded for Set 3 after match is found under Set 2
(E) John Smith Furnitures Inc, Tempe, AZ	500000	7239	Not Considered	

Final Match: In 1993, assignee 25 matched to BR Firm 500000 with Reliability Code B. Assignee matched to a unique establishment.

Cluster 6 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John & Smith Furniture Inc, Tempe, AZ	500000	7240	Set 3	& removed
(E) John Smith Furnitures Inc, Tempe, AZ	500000	7239	Set 3	Furnitures=Furniture

Final Match: In 1993, assignee 25 matched to BR Firm 500000 with Reliability Code C. Assignee not matched to a unique establishment.

Cluster 7 (Blocked on city and state for year 1993):

Assignee (A) /Establishment (E) Name	Assignee Number/MTCHVAR	Employer Identifier	Matches First Under Rule Set	Comments
(A) John Smith Furniture Inc, Tempe, AZ	25			
(E) John & Smiht Furniture Inc, Tempe, AZ	500000	5324	Never Matched	Smiht!=Smith
(E) Jon Smith Furniture Inc, Tempe, AZ	500000	7239	Never Matched	Jon!=John

Final Match: In 1993, assignee 25 is not matched to any BR firm.

APPENDIX II: MERGING FIRM DATA AND PATENT DATA - SOME EXAMPLES

SCENARIO 1: YEAR 1992

Patent Data

Assignee Name	ASSIGNEE	Patent Stock
John Smith Inc	25	12

Firm Data

Headquarters (H)/Subsidiary or Branch	MTCHVAR	ESTAB. ID	Employment
John Smith Inc (H)*	500000	500000001	50
The John Smith Company	500000	500000002	1000
Business Inc	500000	500000003	500
Miami Furniture	500000	500000004	1500

*: Links to Assignee in the NBER Patent Data-BR Bridge

Matched Patent-Firm Data

ASSIGNEE	MTCHVAR	Patent Stock	Total Emp.
25	500000	12	3050

SCENARIO 2: YEAR 1992

Patent Data

Assignee Name	ASSIGNEE	Patent Stock
John Smith Inc	25	12

Firm Data

Headquarters (H)/Subsidiary or Branch	MTCHVAR	ESTAB. ID	Employment
Intelligent Corporation (H)	500000	500000001	0
The John Smith Company	500000	500000002	1000
John Smith Inc*	500000	500000003	500
Miami Furniture	500000	500000004	1500

*: Links to Assignee in the NBER Patent Data-BR Bridge

Matched Patent-Firm Data

ASSIGNEE	MTCHVAR	Patent Stock	Total Emp.
25	500000	12	3000

SCENARIO 3: YEAR 1992

Patent Data

Assignee Name	ASSIGNEE	Patent Stock
John Smith Inc	25	12

Firm Data

Headquarters (H)/Subsidiary or Branch	MTCHVAR	ESTAB. ID	Employment
Intelligent Corporation (H)	500000	500000001	0
The John Smith Company	500000	500000002	1000
John Smith Inc*	500000	500000003	0
Miami Furniture	500000	500000004	1500

*: Links to Assignee in the NBER Patent Data-BR Bridge. Even though John Smith Inc has no employees, its establishments will appear in the BR because other subsidiaries will have to pay payroll tax.

Matched Patent-Firm Data

ASSIGNEE	MTCHVAR	Patent Stock	Total Emp.
25	500000	12	2500

APPENDIX III: TREATMENT OF ACQUISITIONS - SOME EXAMPLES

SCENARIO 1: John Smith Inc is a firm that exists from 1992 to 1994 in the BR. It acquires Patent Corporation in 1993 which was a single establishment firm before the acquisition. John Smith Inc did not patent before the acquisition, and continues to patent under the assignee name "Patent Corporation" after the acquisition. Furthermore, Patent Corporation continues to exist as an establishment.

Firm Data

	Year	MTCHVAR	Employment
John Smith Inc	1992	500000	900
John Smith Inc	1993	500000	1000
John Smith Inc	1994	500000	1100
Patent Corporation	1992	035219867	7
Patent Corporation	1993	500000	9
Patent Corporation	1994	500000	15

Patent Data

Assignee Name	Year	ASSIGNEE	Patent Stock
Patent Corporation	1992	35	10
Patent Corporation	1993	35	12
Patent Corporation	1994	35	15

* John Smith Inc does not patent, and hence, does not exist in the Patent Data.

NBER Patent Data-BR Bridge

Year	ASSIGNEE	MTCHVAR
1992	35	035219867
1993	35	500000
1994	35	500000

Matched Patent-Firm Data (NO INACCURACIES)

MTCHVAR	Year	Patent Stock	Total Emp.
035219867	1992	10	7
500000	1992	0	900
500000	1993	12	1009
500000	1994	15	1115

SCENARIO 2: John Smith Inc is a firm that exists from 1992 to 1994 in the BR. However, John Smith Inc has patented since 1991, and appears in the patent data from that year. It acquires Patent Corporation in 1993, and continues to obtain some patents under "Patent Corporation" and some others under "John Smith Inc" after the acquisition.

Firm Data

	Year	MTCHVAR	Employment
John Smith Inc	1992	500000	900
John Smith Inc	1993	500000	1000
John Smith Inc	1994	500000	1100
Patent Corporation	1992	035219867	7
Patent Corporation	1993	500000	9
Patent Corporation	1994	500000	15

Patent Data

Assignee Name	Year	ASSIGNEE	Patents Applied	Patent Stock
John Smith Inc	1991	25	15	15
John Smith Inc	1992	25	13	28
John Smith Inc	1993	25	20	48
John Smith Inc	1994	25	30	78
Patent Corporation	1992	35	10	10
Patent Corporation	1993	35	2	12
Patent Corporation	1994	35	3	15

NBER Patent Data-BR Bridge

Year	ASSIGNEE	MTCHVAR
1992	25	500000
1993	25	500000
1994	25	500000
1992	35	035219867
1993	35	500000
1994	35	500000

*There is no match in 1991 because John Smith Inc does not exist in the BR that year.

Matched Patent-Firm Data (NO INACCURACIES)

MTCHVAR	Year	Patents Applied	Patent Stock	Total Emp.
035219867	1992	10	10	7
500000	1992	13	28	900
500000	1993	15 (13 +2)	60 (48+12)	1009
500000	1994	33 (30 +3)	93 (78+15)	1115

SCENARIO 3: John Smith Inc is a firm that exists from 1992 to 1994 in the BR. In 1993, it acquires Patent Corporation, which was a single establishment firm before the acquisition. John Smith Inc did not patent before the acquisition, and patents under the assignee name "John Smith Inc" after the acquisition. Patent Corporation is dissolved after the acquisition.

Firm Data

	Year	MTCHVAR	Employment
John Smith Inc	1992	500000	900
John Smith Inc	1993	500000	1000
John Smith Inc	1994	500000	1100
Patent Corporation	1992	035219867	7

Patent Data

Assignee Name	Year	ASSIGNEE	Patent Stock
Patent Corporation	1992	35	10
John Smith Inc	1993	55	1
John Smith Inc	1994	55	3

NBER Patent Data-BR Bridge

Year	ASSIGNEE	MTCHVAR
1992	35	035219867
1993	55	500000
1994	55	500000

Matched Patent-Firm Data (CONTAINS INACCURACIES)

MTCHVAR	Year	Patent Stock	Total Emp.	Corr. Patent Stock
035219867	1992	10	7	10
500000	1992	0	900	0
500000	1993	1	1000	11
500000	1994	3	1100	14

SCENARIO 4: John Smith Inc is a firm that exists from 1992 to 1994 in the BR. It acquires Patent Corporation in 1993 which was a single establishment firm before the acquisition. John Smith Inc did not patent before the acquisition, and patents under the assignee name "John Smith Inc" after the acquisition. Patent Corporation continues to exist after the acquisition but is not involved in patenting activities.

Firm Data

	Year	MTCHVAR	Employment
John Smith Inc	1992	500000	900
John Smith Inc	1993	500000	1000
John Smith Inc	1994	500000	1100
Patent Corporation	1992	035219867	7
Patent Corporation	1993	500000	10
Patent Corporation	1994	500000	12

Patent Data

Assignee Name	Year	ASSIGNEE	Patents Applied	Patent Stock
Patent Corporation	1992	35	10	10
Patent Corporation	1992	35	0	10
Patent Corporation	1992	35	0	10
John Smith Inc	1993	55	1	1
John Smith Inc	1994	55	2	3

NBER Patent Data-BR Bridge

Year	ASSIGNEE	MTCHVAR
1992	35	035219867
1993	35	500000
1994	35	500000
1993	55	500000
1994	55	500000

Matched Patent-Firm Data (NO INACCURACIES)

MTCHVAR	Year	Patents Applied	Patent Stock	Total Emp.
035219867	1992	10	10	7
500000	1992	0	0	900
500000	1993	1	11	1000
500000	1994	2	13	1100

SCENARIO 5: John Smith Inc is a firm that exists from 1992 to 1994 in the BR. In 1993, it acquires Patent Corporation which was a single establishment firm before the acquisition. John Smith Inc did not patent before the acquisition. After the acquisition, John Smith Inc changes the name of Patent Corporation to John Smith Technologies, which continues to exist as an establishment in the BR. In the patent data, patent Corporation and John Smith technologies are treated as different assignees.

Firm Data

	Year	MTCHVAR	Employment
John Smith Inc	1992	500000	900
John Smith Inc	1993	500000	1000
John Smith Inc	1994	500000	1100
Patent Corporation	1992	035219867	7
John Smith Technologies	1993	500000	9
John Smith Technologies	1994	500000	15

Patent Data

Assignee Name	Year	ASSIGNEE	Patent Stock
Patent Corporation	1992	35	10
John Smith Technologies	1993	5	2
John Smith Technologies	1994	5	5

NBER Patent Data-BR Bridge

Year	ASSIGNEE	MTCHVAR
1992	35	035219867
1993	5	500000
1994	5	500000

Matched Patent-Firm Data (CONTAINS INACCURACIES)

MTCHVAR	Year	Patent Stock	Total Emp.	Corr. Patent Stock
035219867	1992	10	7	10
500000	1992	0	900	0
500000	1993	2	1000	12
500000	1994	5	1100	17