Improving Patent Assignee-Firm Bridge with Web Search Results*

Yuheng Ding† Karam Jo‡ Seula Kim§

November 20, 2023

[Click Here for the Latest Version]

Abstract

This paper constructs a patent assignee-firm longitudinal bridge between U.S. patent assignees and firms using firm-level administrative data from the U.S. Census Bureau. We match granted patents applied between 1976 and 2016 to the U.S. firms recorded in the Longitudinal Business Database (LBD) in the Census Bureau. Building on existing algorithms in the literature, we first use the assignee name, address (state and city), and year information to link the two datasets. We then introduce a novel search-aided algorithm that significantly improves the matching results by 7% and 2.9% at the patent and the assignee level, respectively. Overall, we are able to match 88.2% and 80.1% of all U.S. patents and assignees respectively. We contribute to the existing literature by 1) improving the match rates and quality with the web search-aided algorithm, and 2) providing the longest and longitudinally consistent crosswalk between patent assignees and LBD firms.

Keywords: Innovation, Patent, Patent-firm Concordance, Linked Administrative Data

*Any views expressed are those of the authors and not those of the Korea Development Institute or the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this information product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2095. (CBDRB-FY22-P2095-R9872)

†World Bank. Email: yding4@worldbank.org

‡Korea Development Institute. Email:karamjo@gmail.com

§Princeton University. Email: sk6285@princeton.edu

1 Introduction

It is well known that firm innovation is a major source of creative destruction and productivity growth in the economy. The development of new technologies and products by innovative firms plays a substantial role in not just technology advancement level but also job creation and output growth in the economy. Thus, many researchers have sought to measure and understand innovation activities and their impact on macroeconomic outcomes.

However, measuring firm innovation has been challenging to researchers for various reasons. One of them is the lack of agreement on what constitutes innovation, and another reason is related to measurement issues. Broadly speaking, there are two ways to measure innovation. One is by inputs to innovation activities, such as using R&D expenditure recorded in data. Another way is to track outputs of innovation, such as patents, products, knowledge, new process, or methodology. As well illustrated in previous literature (Balasubramanian and Sivadasan, 2010, 2011; Graham et al., 2018), data on firm innovation have limitation in both dimensions, given that the outputs are hard to quantify (and used to be constrained by the lack of large-scaled data) and the inputs are imperfectly measured and mostly skewed to large firms.

Patents have long been acknowledged as a rich source of data for studying innovation and technology change. Although the United States Patent and Trademark Office (USPTO) make available their patent data to the public, linking the granted patents to the owning firms is nontrivial. USPTO does not keep track of the same assignees over time by giving them unique identifiers, and there is no consistent format for inputting assignee names and addresses. Thus, the raw patent data from the USPTO suffer from a longitudinal inconsistency problem that arises mainly from the misspelling of assignee names.

Pioneered by Hall et al. (2001), there have been a number of prior efforts to overcome this issue and link patent data from the USPTO to firm-level data to track patenting behavior and measure innovation at the firm level. By standardizing assignee names and

using their addresses, Hall et al. (2001) construct the NBER Patent Data, which compile data on all utility patents granted by the USPTO and linked to Compustat. Building on this, Kerr and Fu (2008), Balasubramanian and Sivadasan (2010), and Balasubramanian and Sivadasan (2011) link patent assignees to administrative firm-level data from the U.S. Census Bureau. Graham et al. (2018) and Dreisigmeyer et al. (2018) further extend these efforts by exploiting not just the business assignee information but also inventor information, and using a triangulation methodology across them. Nevertheless, these existing crosswalks have covered either only years before or after 2000.[1] This discontinuity imposes difficulties for researchers who are interested in tracing the dynamics of firm innovation activities in a longer time horizon. In particular, by using different crosswalks for years before and after 2000, it is likely to introduce sample selection bias imposed by different linking methodologies. As the linking methodologies used are different, each bridge shows different match rates between the population of patents applied by the U.S. assignees and firms in the U.S. Census datasets, ranging from around 72% to over 90% at the patent level.

To fill this gap and improve matching quality, we build on earlier approaches by introducing the internet search-aided algorithm in Autor et al. (2020), which utilizes machine-learning capacities of a web search engine. We construct a concordance between the USPTO assignees and firms in the Business Register (BR) and the Longitudinal Business Database (LBD) of the Census Bureau.

Overall, we are able to match 88.2% and 80.1% of all U.S. patents and assignees respectively between 1976 and 2016.[2] In particular, the search-aided algorithm improves the match rates by 7% and 2.9% at the patent and assignee level, respectively. More

[1]The first version of the NBER Patent Data covers the U.S. patents granted between 1963 and 1999, and later it was extended to 2006. Balasubramanian and Sivadasan (2010) and Balasubramanian and Sivadasan (2011) use the first version of the NBER Patent Data, and Kerr and Fu (2008) use the extended version of the NBER Patent Data up to 2002. If we consider the well-known right truncation issue of the USPTO patent data that comes from a lag that occurs in a granting process, their bridges can be used for analysis up to 1999 at most. Graham et al. (2018) and Dreisigmeyer et al. (2018) bridges start from 2000 as the Longitudinal Employer-Household Dynamics (LEHD) they use to grab inventor information is only available from 2000 for a majority of states.

[2]Note that we achieve these match rates even without manually matching assignees that our fully automated algorithm may have failed to capture.

importantly, our methodology significantly extends the time period and screens all U.S. patent assignees between 1976 and 2016. Thus, to the best of our knowledge, our crosswalk is by far the only one that covers this long time horizon between the USPTO and the Census with a consistent linking methodology. Thus, we expect to bring broad benefits to researchers studying dynamics and evolution of innovation and technology within firms. In particular, our bridge will be useful to researchers who are interested in entrepreneurship, technological knowledge and innovation activities by small or young firms that are not recorded in publicly available data such as Compustat. Furthermore, the bridge will allow deeper analysis of firm innovation over a long period of time.

This paper aims to provide details of our methodology used to develop the new bridge and to facilitate the use of this bridge by researchers. The rest of the paper proceeds as follows. Section 2 illustrates various data sources we use to build our crosswalk. Section 3 describes the matching methodology. Section 4 presents match results. Section 5 provides the benefits of the bridge along with various examples illustrating the practical application of the bridge. Lastly, Section 6 concludes by addressing limitations in the matching methodology and suggests areas for potential improvement in future research.

2 Data Sources

We use three datasets to build a longitudinal bridge between patent assignees and firms in the U.S. Census Bureau datasets: the USPTO PatentsView database, the Business Register (BR), and the Longitudinal Business Database (LBD). The last two datasets are administrative data hosted by the U.S. Census Bureau.

2.1 USPTO PatentsView Database

The USPTO PatentsView tracks all patents granted by the USPTO from 1976 onward.[3] This database contains detailed information for granted patents, including application and grant dates, technology class, patent citation information, and the name and address of patent assignees.

The raw USPTO data contain a comprehensive set of information about all types of patents, including utility, design, plant, and reissued patents. Of these, utility patent accounts for over 90% of the granted patents in the U.S., which covers the creation of new or improved products, machines, and processes. Furthermore, the raw data classify patent assignees by the following categories: 1) Unassigned; 2) U.S. company or corporation; 3) Foreign company or corporation; 4) U.S. individual; 5) Foreign individual; 6) U.S federal government; 7) Foreign government; 8) U.S. county government; and 9) U.S. state government.

We use the USPTO raw data version downloaded on December 29, 2020, and extract patents applied from 1976 through 2016 to capture firm innovation activities in this period. This contains over 6 million patents in total. We use the application and grant year of each patent, and the name and address information (state and city) of assignees associated with a given patent in our matching algorithm.

Table 1 to 4 provide descriptive summary statistics of the raw USPTO PatentsView data. Specifically, Table 1 shows the share of patents applied in 1976-2016 by patent type.[4] There are more than six million patents, among which 92.56% are utility patents. Table 2 shows that around 48.86% of patents are assigned to U.S. company/corporation, while 49.24% are assigned to foreign company/corporation. As shown in Table 3 and 4, the number of applied and granted patents show upward trends over our sample period, 1976-2016. Amongst those patents, our main interests lie in patents applied to U.S. company and corporation from 1976 through 2016. In later section, we present

[3] See details from https://patentsview.org/download/data-download-tables.

[4] See more descriptions from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/patdesc.htm

match rates for these patents.

2.2 Data from the U.S. Census Bureau

2.2.1 The Business Register (BR)

The BR (previously referred to as the Standard Statistical Establishment Listing or SSEL) is a comprehensive database of the U.S. business establishments with paid employees, which is a core source of longitudinal business demographics and characteristics about establishments linked to the Longitudinal Business Database (LBD). The dataset contains the establishment-level information such as establishment identifiers, name, address, whether the establishment belongs to single-unit or multi-unit firms, and parent firm identifier associated with each establishment. DeSalvo et al. (2016) contain detailed information about the BR.

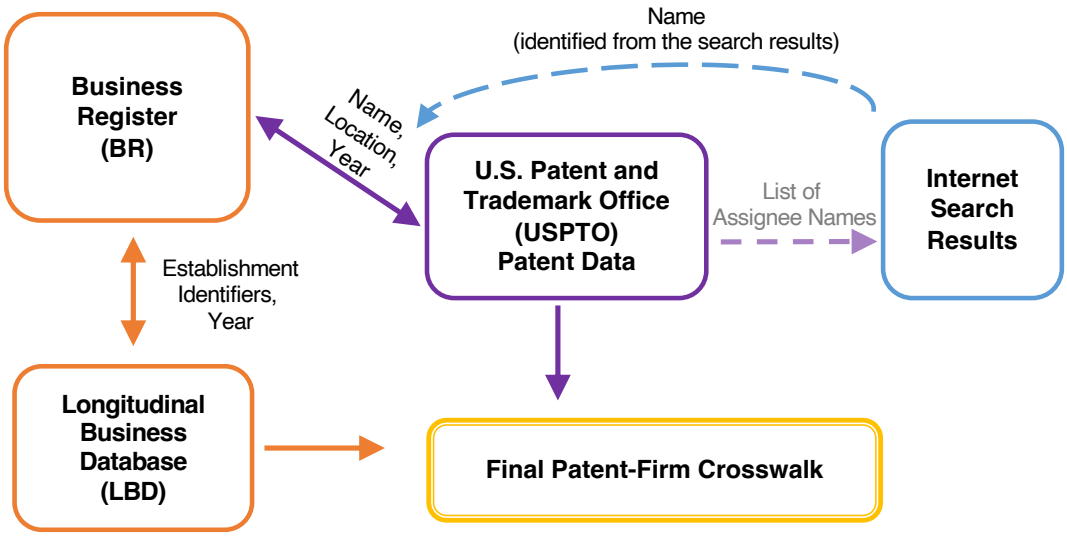
We use the establishment-level name and address (state and city) information from the BR to identify all associated establishments of the patent assignees in the USPTO database.

2.2.2 The Longitudinal Business Database (LBD)

The LBD tracks the universe of private non-farm establishments and firms with at least one paid employee in the U.S. from 1976 and onward. The data provide detailed information about employment, payroll, industry codes, establishment and firm identifiers, employer identification numbers, business name, and location. With the LBD, we can capture firms and establishments that enter or exit each year, along with firm age, defined as the age of the oldest establishment. Jarmin and Miranda (2002) and Chow et al. (2021) present more information about the LBD.

In the LBD, establishments operated by the same entity, identified through the Economic Census and the Company Organization Survey, are grouped under a common firm identifier. We aggregate the establishment-level information into firm-level obser-

Figure 1: Diagram of Patent-Firm Matching and Crosswalk Creation



vations using these firm identifiers.

3 Linking Methodology

As briefly explained before, USPTO does not keep track of a given assignee over time by assigning it a unique identifier. Also, assignee name and address information contains misspelling and abbreviation issues. Our goal is to overcome these issues and link patents in the USPTO PatentsView database to the LBD by constructing a crosswalk between patent assignees and LBD firms in 1976-2016. We treat assignees in the USPTO data as the same entity if they have identical standardized names and location information at the state and city level in a given year (using either patent application or grant year). Following pre-existing literature, we construct and use two standardized assignee names to build the crosswalk. Within each iteration for patent bridge construction, we take the following steps to link USPTO patents to LBD by using strict and fuzzy name matching enhanced by the search-aided approach adopted from Autor et al. (2020). Figure 1 provides an overview of our matching methodology.

First, we use strict and fuzzy name matching techniques with geographical blocking variables to link USPTO patent assignees to business establishments in the BR. We cannot match patent assignees to firms directly through name matching since the BR only contains names at the establishment level but not at the firm level. By the end of this step, we find all business establishments that could possibly be under a corporate umbrella of the firm assignee. Then, we use existing establishment identifiers to find the most reliable LBD firm identifier by comparing the quality of fuzzy name matching results among all establishments that are linked to a patent assignee.

One challenge of integrating the USPTO patent data with the Census data is potential differences in timing between these. Since patents usually take years to go through a whole examination procedure, the assignee information including name and location could vary across application and grant years. Also, firm identifiers in the current LBD are not perfectly time consistent, and information in the BR might not be updated in a timely manner, where the information is the most accurate in the years the Economic Census is conducted (years ending in '2' and '7').

To mitigate these issues and to take into account the fact that application year is closer to the time when firms undertake innovation, we first use patent application year as our reference year, and we use patent grant year to find matches only if there is no match found based on the application years. Moreover, we implement a +3/-3 year time window to improve match rates in cases where only either side of the data has updated information while the other has not.

After the fuzzy name matching, we further implement the internet search-aided algorithm following Autor et al. (2020) to improve our match rates and quality. In essence, the internet search-aided algorithm leverages the machine-learning capacities of the web search engine to identify patent assignees that would have been treated as the same entity if their names were not misspelled. Combined with the most reliable fuzzy name matching results from the previous steps, the internet search-aided algorithm helps us improve the match rates by linking unmatched patent assignees to the same LBD firms as identified in the previous fuzzy name matching procedure. We describe

each of these steps in more detail in the following sections.

3.1 Name Standardization

Before getting into the matching process, we standardize firm names and produce stem names for the USPTO assignees and the BR establishments. For the name standardization and stem name production, we follow the same process as in the NBER Patent Data Project (NBER PDP).[5]

3.2 Name Matching

3.2.1 Match Patent Assignees to BR Establishments and LBD Firms

As the first step of our matching procedure, we match patent assignment information in the USPTO to the BR by using the SAS DQMatch fuzzy matching procedure and using available geographic information as blocking variables. Moreover, to allow potential timing mismatches between the USPTO patent data and the BR data, we use a +3/-3 year window around the reference year. We use the application year and grant year of patents as two alternative years of reference to match patent assignees to BR establishments. Within each reference year loop, the same matching passes are implemented, as shown in Table 5, to match patent assignees with the BR establishments within a +3/-3 year window centered around the reference year.

Specifically, we start with the most restrictive criteria by using strict name, state, and city, e.g., Model A1 in Table 5, to match patent assignees to the BR establishments within the +3/-3 year window. Then, for the unmatched patent assignees, we move on to the next Model A2 and allow "fuzziness" in the name of cities of patent assignees and BR establishments. Again, we keep the unmatched patent assignees from the previous model A2 and match them to the BR establishments by using the SAS DQMatch fuzzy name matching procedure blocking on state and strict city names, and so on for all

[5] See more details from https://sites.google.com/site/patentdataproject/Home.

models in Table 5.

We run the same matching passes and use the patent application and grant years as two alternative reference years. Thus, there are two concordance tables generated by this step, one is the assignee-BR matches based on the patent application year, and the other is based on the patent grant year. In each concordance table, the data contain the assignee name, state, city, the BR year (along with the associated time window) when an establishment is matched to the patent assignee, and either the application or grant year from the USPTO, along with information on the matched establishments from the BR. In other words, this concordance table provides a crosswalk between patent assignees in a given reference year to all matched BR establishments within the +3/-3 year window. It is noteworthy that a patent assignee could be matched to multiple BR establishments if these are establishments of the same firm located in the same city and state. All the matched BR establishments, however, are identified by the same and the most reliable model in Table 5 for a given patent assignee-year.

Next, we use existing establishment identifiers from the Census to link all patent assignee-BR establishment pairs from the previous step to the LBD, and extract firm identifiers. To be specific, we use the BR establishment identifier “cfn" (before 2002) and “empunit_id" (after 2002) to match “estabid" and “estabid_rorg" in the 2018 version redesigned LBD (Chow et al., 2021) to find the associated firm identifier “lbdfid". Since a given assignee could be, in principle, matched to multiple BR establishments that belong to different firms, this can leave us with multiple firm identifiers being matched to a given patent assignee.

To find the most reliable firm identifier for a given patent assignee, we then calculate the Jaro-Winkler similarity between the patent assignee name and all of the BR establishment names linked to it. Thus, for each pair of assignee name-state-city-firm identifier (lbdfid), we keep the one with the highest Jaro-Winkler similarity score. If there are ties, i.e., more than one record having the same highest Jaro-Winkler score, we randomly select one of them and drop the rest. This ensures that a given patent assignee in the same reference year is matched to a unique LBD firm. And it gives us the

firm-level concordance between USPTO patent assignees and LBD firms in all reference years between 1976 and 2016.

3.2.2 Linking Patent ID to LBD Firms

Since our ultimate goal is to link USPTO patents to LBD firms, we use the assignee-firm concordance table from the previous step to link all patents of an assignee to the matched LBD firm. To be specific, all patents having the same assignee name, city, and state in the same reference year are linked to the same LBD firm. As mentioned before, however, we use both the application year and the grant year of patents as two alternative years of reference to match patent assignees to LBD firms in the previous step. Thus, there could potentially be at most two possible matches for a given patent, one matched by the application year and the other by the grant year, since we have already sorted out the best match and chosen a unique firm identifier associated with the highest Jaro-Winkler similarity score. Nevertheless, a given patent could be matched to different firm identifiers depending on which reference year we are using.

In those cases where we have two inconsistent matches, we choose the more reliable one by comparing the quality of the matching models, as in Table 5. If the two matches are found based on the same criteria in Table 5, we then compare the year gaps between the matched results and the reference year. Table 6 provides the preference ordering we implement. In general, we consider matches having no year gaps the most reliable, and then years preceding the reference year, e.g., years t-1, t-2, and t-3, followed by years after the reference year, e.g., years t+1, t+2, t+3. For instance, suppose a patent is matched to Firm A by application year but Firm B by grant year based on the same model in Table 5. If Firm A is found in the same year as the patent application year while Firm B is not found in the same year as the grant year, we trust the result of Firm A more and drop Firm B. If there is still a tie, e.g., having Firm A and B found in the same LBD year as the patent application year and grant year accordingly, we prioritize the result based on the patent application year, e.g., Firm A. over the one found in the

patent grant year.

As explained previously, we use both the application year and the grant year to improve the overall match rates by mitigating potential issues that arise from the time inconsistency of the LBD firm identifiers across time. Importantly, we prefer matching results based on the application year due to the fact that the point in time when innovation activities occur is better measured by the time of patent application rather than the time when a patent is granted—the granting process takes three to five years on average, and sometimes more than ten years. In addition to the timing difference between the USPTO and the Census data, which might lead to a higher rate of false negative results, changes in the assignee firm's organizational structure could potentially result in changes in firm identifier between the year of patent application and the year when the patent is granted. Such cases include merger and acquisition, a transition between single-unit and multi-unit firms, as discussed in more detail in Chow et al. (2021).

Thus, using grant years only might also lead to false positives or create difficulties in identifying the correct firm identifier at the time of innovation activities. By using both application year and grant year while prioritizing results based on the application year, our approach mitigates issues caused by the time-inconsistent firm identifier while ensuring that the false negative rate is the same, if not lower, as matching results based on similar approaches that use grant year only as the reference year.

While we have run the matching models from A1 to G4 in Table 5, extra caution should be applied when using matching results based on models after D as the geographic information is, in general, not fully matched. Thus, we only keep assignee-firm matches identified by Model A-D to construct a reference table that we used in the next step. For all of the remaining matches after Model D, as well as all the unmatched assignees and patents, we implement the following internet search-aided approach to minimize the number of false negative results.

3.3 The Internet Search-aided Algorithm

The self-reported firm assignee names entered on patents often contain unusual abbreviations or misspellings, which causes a long-standing challenge and imposes limitations on linking the patent database with other firm-level data products. Although our previous matching processes, including name standardization and fuzzy name matching through the SAS DQMatch procedure, can handle many misspelling issues, there are still complicated cases that cannot be adequately coped with by those techniques. Thus, we follow Autor et al. (2020) and leverage the machine-learning capacities of the internet search engine to improve our match results further.

The methodology works as follows. Suppose that we would like to make a bridge between Data A and B through the web search results. We first enter the assignee names on Data A, collect the top five search results, and do the same for Data B. Next, we compare the search results across the two databases and consider firms to be identical if they share at least a certain sufficient amount of the same search results out of the top five results. For "International Business Machines" and "IBM," for example, an internet search engine will suggest IBM.com and IBM's Wikipedia page as its top search results for both firm names.

Autor et al. (2020) match the USPTO patents to Compustat firms. Since both datasets are publicly available, they use the internet search engine to collect top search results for all firm names in both datasets and directly compare the search results across the two datasets. In building the crosswalk between patents and LBD firms, however, we are unable to search firm names in the BR since the Census administrative data can only be accessed at the Census Research Data Center (RDC), where connection to the internet is not allowed.

Therefore, we are only able to extract top search results for USPTO patent assignee names outside of the RDC. Specifically, we put every patent assignee name into the Google.com search engine, collect the URLs of the top five search results, and identify any given pair of patent assignees as the same firm if they share at least two identi-

cal search results. In essence, we utilize the internet search results to unify assignee names in the USPTO patent data and then create a concordance table between original patent assignee names and the unified assignee names determined by the internet search results.

Next, we bring this concordance table to the RDC and combine it with the previously constructed patent-firm crosswalk to reduce false negative results. To do so, we rely on previous matches identified by Model A-D in Table 5 to construct a reference bridge between patent assignees and the BR establishments. Then, we combine this reference bridge with the concordance table of patent assignee names generated by the internet search results to find BR establishment names for those unmatched patent assignees as well as those with low matching quality, e.g., those matched by Model E-G. In other words, this approach helps us find additional linkages of patent assignee names to BR establishment names that have already been matched to the patent assignees whose names are spelled differently but should have been identified as the same entity names based on the internet search results.

Then, we apply the same criteria as those in Section 3.2 to the new matches. We rank the matched BR establishments following Table 5 as before and keep those by the most reliable model only. Note that the original names of the newly matched USPTO assignees won't be matched to those of the matched BR establishments through the standard name matching process. The reason is that these are first linked to the USPTO assignees in the reference bridge through the web search results and then to the BR establishments through the reference bridge.

Thus, we use the matched USPTO assignee names in the reference bridge for this procedure instead of the original names of the newly matched USPTO assignees. For the city and state information, we use the ones initially attached to the newly matched USPTO assignees. Moreover, we impose a restriction on year windows by dropping any matched results if neither their application nor grant year is the same as the BR year of the matched establishment(s). The remaining steps in Section 3.2 are applied as before.

3.4 Stem Name Matching

Although the internet search-aided algorithm helps us improve the matching by increasing the number of matches, there are still other unmatched patents (or assignees) for which we would like to find matches. To do this, we use the unmatched set of the USPTO database and identify the unique pair of the patent assignee stem name, city, state, and either application or grant year. We merge this with the BR establishment as in the first step (Section 3.2), but now by the strict stem names. And we sort out the matches by the model ordering in Table 7, which is only by the address information given that we have already gone through the strict name matching with stem names. The remaining parts follow the same as before as in Section 3.2. Here again, from the final set of matches, we only use those from model AA-DD to be consistent. In other words, we only include the matches through the model AA-DD to the previous set of matches, and treat the rest as unmatched.

3.5 The 2nd Search-aided Approach

Lastly, we apply the search-aided approach again by incorporating the results from the stem name matching. As before, we start with the web search results for the USPTO assignees from Section 3.3, collect still unmatched assignees, and merge them with the reference bridge that includes the stem name matching results (with models AA-DD) in Section 3.4. This constructs a search-identified bridge between the USPTO assignee names and the BR establishment names.

As we did before in Section 3.3, we rank the matches by the model ordering in Table 5 if the match comes from the reference bridge through the initial standardized name matches, and in Table 7 if it is from the reference model associated with the stem name matches. Again, we evaluate these by using the state and city information of the new assignees identified through the search bridge in this step. And then, the rest procedure follows as before. This finalizes the matched set of data.

4 Match Results

In this section, we present the match results of utility patents assigned to U.S. non-government organizations. However, the current summary statistics of the match results are limited due to disclosure risk. We plan to provide more comprehensive descriptions of the matching results in an internal technical memo accessible to Census and RDC researchers.

Table 8 shows the overall match rates of U.S. patents and assignees from 1976 to 2016 by aggregate model types. Specifically, we classify our match models into the following categories: i) standardized name matches - "Model STD D" that are based on model A-D in Table 5); ii) stem name matches - "STEM D+" that are based on model AA-DD in Table 7; iii) "ISRA" models based on the search-aid algorithm; iv) remaining matches based on model E-G and EE-FF in Table 5 and 7 accordingly; and v) no match results. The first column shows the match rates for U.S. patents, and the second column presents the match rates for the U.S. assignees.

Overall, the match rates of the U.S. patents and assignees are 88.2% at the patent level and 80.1% at the assignee level. Even with the most reliable models (STD D+, STEM D+, and ISRA), the match rates are 83.7% at the patent level and 71.2% at the assignee level. It is noteworthy that more than half of the matches are based on the standardized name matching models (model A-D in Table 5). The STD D+ models are able to match 62% of patents and 55.5% of assignees. Moreover, name-location matching based on STEM name, e.g., STEM D+, is able to improve the patent- and assignee-level results further by 14.8% and 12.8%, respectively.

Most importantly, the internet search-aided algorithm significantly improves the match rates by 7% at the patent level and 2.9% at the assignee level. This improvement accounts for 8.5% of the total patent-level matches and 4.1% of the entire assignee-level matches. We consider matching results based on model STD D+, STEM D+ and ISRA the most reliable ones as the probability of false positive matches is low [6].

[6] Specific numbers cannot be disclosed by the Census Bureau at this stage

As mentioned, we further relax the matching criteria by removing geographic blocking variables, e.g., state and city, to minimize the risk of false negatives. As shown in the last row of Table 8, we find that model E-G (for the standardized name matching) and model EE-FF (for the stem name matching) further improve the match rates by 4.5% and 8.9% at the patent and assignee level, respectively. Nevertheless, these matching results might incorporate false positive linkages more. More information on false positive rates will be provided in a future draft that can be circulated internally to researchers having access to the Federal Statistics Research Data Centers.

5 Benefits and Real-world Applications of the Bridge

In this section, we provide the advantages of the bridge and several practical examples of its application in research.

As detailed in the preceding section, our bridge significantly improves the quality of the match between patent data and firm-level administrative data. Additionally, the uniqueness of our concordance arises from its extended sample period and inclusion of non-public firms in its coverage. This is essential for studies focusing on tracing firm innovation (especially for small or younger firms) over an extended period of time. Our crosswalk is thus essential for a broad area of economic research, including topics related to innovation, technology evolution, firm strategy, and economic growth.

To this end, we contribute to existing efforts of building patent assignee-firm crosswalks from the following two aspects. First, the internet search-aided algorithm enhances the quality of matches between patents and U.S. firms by a non-negligible fraction. And we find that the internet search-aided matching algorithm substantially increases the match rates at both patent and assignee-level. Moreover, this implies the validity and feasibility of implementing the internet search-aided matching algorithm to construct similar crosswalks based on name matching algorithms.

Second, our match results remarkably extend the sample period covering all USPTO granted patents during 1976-2016. To the best of our knowledge, this is the longest

longitudinal patent assignee-firm bridge using the Census data, which is conducive to researchers wanting to study firm innovation over such an extensive period of time.

There are several examples of the real-world application of the bridge. One illustrative case involves examining the impact of competition on firm innovation and business dynamism, leveraging China's WTO accession in 2001 as an exogenous competition shock, as demonstrated in studies such as Jo (2019) and Jo and Kim (2021). In this context, our bridge is particularly useful in three key dimensions.

First, the bridge encompasses both pre- and post-2000 periods and facilitates the identification of the causal effect of the Chinese competition. This is achieved by enabling the authors to utilize the rise of China in the U.S. markets after China's WTO accession in 2001 as a quasi-experimental increase in competition induced by foreign firms. Furthermore, the authors use a Difference-in-Difference (DD) specification to identify the China competitive pressure shock on U.S. firm innovation as in Pierce and Schott (2016). Given that, our bridge further allows the authors to test the parallel pre-trends assumption – the crucial identifying assumption for the DD model – by spanning periods even before the 1990s. Moreover, the bridge enables the authors to investigate the effect of Chinese competition on firm entry, young firm activities, and business dynamism by incorporating non-public firms, particularly those that are small and/or young and not recorded in publicly available datasets.

Another example includes studies tracing the evolution and transformation of firm innovation over time through the interactions between heterogeneous firms in the U.S. economy. With technological advancements, innovation tends to necessitate a larger knowledge base, increasingly achieved by large firms comprised of specialized experts (Jones, 2009). Examining these shifts in firm innovation and knowledge complexity over a long horizon of period or across different groups of firms, such as size, age, or cohorts, and assessing their implications for overall business dynamism and economic growth presents a set of compelling research questions that can leverage the capabilities of our bridge. For instance, we can use our bridge to investigate how the trends in large, established firms' innovation activities (such as mega firms producing more novel

patents in 2000s as in Braguinsky et al. (2023)) are interacted with the changes in small or young firms' innovation activities and business dynamism over time. Aside these examples, there are various other scenarios where our bridge can be effectively applied.

6 Conclusion and Future Work

In this paper, we construct a longitudinally consistent linkage between the U.S. patent assignees and firms recorded in the U.S. Census administrative data between 1976-2016. Our method differs from previous patent matching efforts by introducing an internet search-aided algorithm with an extended time horizon. Indeed, existing bridges between patent assignees and firms are either only for publicly available firm-level data or based on a standard name-location matching process with a shorter time period. Utilizing an internet search-aid matching algorithm, we improve the matching quality of existing assignee-firm crosswalks in the literature and significantly extend the time horizon of similar crosswalks.

Although the bridge is expected to bring a broader set of benefits to researchers compared to the existing crosswalks, there is still room for improvement. First, The current matching procedures do not include the process of manual matching. Therefore, the current matches can be further improved both qualitatively and quantitatively if we incorporate match results obtained by manual matching. Indeed, we plan to conduct manual matching for the unmatched patents to minimize the risk of false negatives as well as to screen the matched results to remove any false positive results.

Also, false positive results are inevitable even from the most reliable model in Table 5. In rare circumstances, establishments having the same name, state, and city in a given year are associated with different firm identifiers. There is no plausible way, however, to figure out which of those firms is the correct assignee given the set of information we have in data. False positive results like these could be more pronounced for models with missing or inconsistent location information, such as models B, C, and D.

In order to have a better understanding of the likelihood of false positives, we have estimated the false positive rates in the Census BR for different model types. However, we are unable to report details in this draft due to Census disclosure requirements. If possible, we would like to present more detailed match information and statistics in future versions of the paper. We are also working on technical notes that are accessible to researchers who have access to the Federal Statistical Research Data Centers.

References

D. Autor, D. Dorn, G. H. Hanson, G. Pisano, and P. Shu. Foreign competition and domestic innovation: Evidence from us patents. *American Economic Review: Insights*, 2(3):357–74, 2020.

N. Balasubramanian and J. Sivadasan. Nber patent data-br bridge: User guide and technical documentation. *US Census Bureau Center for Economic Studies Paper No. CES-WP-10-36*, 2010.

N. Balasubramanian and J. Sivadasan. What happens when firms patent? new evidence from us economic census data. *The Review of Economics and Statistics*, 93(1):126–146, 2011.

S. Braguinsky, J. Choi, Y. Ding, K. Jo, and S. Kim. Mega firms and recent trends in the us innovation: Empirical evidence from the us patent data. Technical report, National Bureau of Economic Research, 2023.

M. C. Chow, T. C. Fort, C. Goetz, N. Goldschlag, J. Lawrence, E. R. Perlman, M. Stinson, and T. K. White. Redesigning the longitudinal business database. Technical report, National Bureau of Economic Research, 2021.

B. DeSalvo, F. F. Limehouse, and S. D. Klimek. Documenting the business register and related economic business data. *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-17*, 2016.

D. Dreisigmeyer, N. Goldschlag, M. Krylova, W. Ouyang, E. Perlman, et al. Building a better bridge: Improving patent assignee-firm links. Technical report, Center for Economic Studies, US Census Bureau, 2018.

S. J. Graham, C. Grim, T. Islam, A. C. Marco, and J. Miranda. Business dynamics of innovating firms: Linking us patents with administrative data on workers and firms. *Journal of Economics & Management Strategy*, 27(3):372–402, 2018.

B. H. Hall, A. B. Jaffe, and M. Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools, 2001.

R. S. Jarmin and J. Miranda. The longitudinal business database. *Available at SSRN 2128793*, 2002.

K. Jo. Defensive innovation and firm growth in the us: Impact of international trade. Technical report, Working Paper, 2019.

K. Jo and S. Kim. Competition, firm innovation, and growth under imperfect technology spillovers. *Available at SSRN 3850146*, 2021.

B. F. Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

W. R. Kerr and S. Fu. The survey of industrial r&d—patent database link project. *The Journal of Technology Transfer*, 33(2):173–186, 2008.

J. R. Pierce and P. K. Schott. The surprisingly swift decline of us manufacturing employment. *American Economic Review*, 106(7):1632–1662, 2016.

Tables

Table 1: Frequency of Patent Type

Patent Category	Counts	Percent (%)
Utility	5,846,531	92.56
Design	433,110	6.86
Plant	20,424	0.32
Reissue	16,350	0.26
Defensive Publication	210	0
TVPP	3	0
Total	6,316,628	100

Notes: This table shows the distribution of patent categories from the USPTO raw data (December 29, 2020 version). “Utility" is patents issued for the invention of a new and useful process, machine, manufacture, or composition of matter, or a new and useful improvement; “Design" means design patents that are issued for a new, original, and ornamental design embodied in or applied to an article of manufacture; “Plant" refers to plant patents issued for a new and distinct, invented or discovered asexually reproduced plant including cultivated sports, mutants, hybrids, and newly found seedlings, other than a tuber propagated plant or a plant found in an uncultivated state; “Reissue" is those re-issued to correct an error in an already issued utility, design, or plant patent; “Defensive Publication (DEF)" is patents issued instead of a regular utility, design, or plant patent; and lastly,“TVPP" refers to Trial Voluntary Protest Program (TVPP) patents.

Table 2: Frequency of Assignee Type

Assignee Category	Counts	Percent (%)
Unassigned	8	0
U.S. Company/Corporation	3,086,203	48.86
Foreign Company/Corporation	3,110,310	49.24
U.S. Individual	37,244	0.59
Foreign Individual	27,905	0.44
U.S. Federal Government	40,418	0.64
Foreign Government	14,258	0.23
U.S. County Government	25	0
U.S. State Government	257	0
Total	6,316,628	100

Notes: This table shows the distribution of assignee categories from the USPTO raw data (December 29, 2020 version).

Table 3: Frequency of Patents by Application Year

Application Year	Counts	Percent (%)
1976	55,329	0.88
1977	55,858	0.88
1978	55,761	0.88
1979	56,190	0.89
1980	57,946	0.92
1981	56,978	0.90
1982	59,351	0.94
1983	56,485	0.89
1984	61,310	0.97
1985	65,637	1.04
1986	68,306	1.08
1987	73,727	1.17
1988	81,678	1.29
1989	87,499	1.39
1990	90,505	1.43
1991	92,287	1.46
1992	96,567	1.53
1993	100,029	1.58
1994	115,024	1.82
1995	137,209	2.17
1996	138,036	2.19
1997	162,594	2.57
1998	163,376	2.59
1999	176,231	2.79
2000	194,084	3.07
2001	209,310	3.31
2002	210,112	3.33
2003	203,356	3.22
2004	206,905	3.28
2005	213,632	3.38
2006	220,603	3.49
2007	229,420	3.63
2008	232,892	3.69
2009	223,835	3.54
2010	238,363	3.77
2011	259,096	4.10
2012	283,400	4.49
2013	300,564	4.76
2014	308,040	4.88
2015	314,356	4.98
2016	304,747	4.82
Total	6,316,628	100

Table 4: Frequency of Patents by Grant Year

Grant Year	Counts	Percent (%)
1976	815	0.01
1977	23,553	0.37
1978	49,691	0.79
1979	40,202	0.64
1980	51,098	0.81
1981	55,947	0.89
1982	50,470	0.80
1983	50,833	0.80
1984	60,083	0.95
1985	64,400	1.02
1986	63,486	1.01
1987	74,361	1.18
1988	70,276	1.11
1989	85,185	1.35
1990	81,727	1.29
1991	88,201	1.40
1992	89,969	1.42
1993	92,323	1.46
1994	95,473	1.51
1995	95,417	1.51
1996	103,135	1.63
1997	106,628	1.69
1998	140,798	2.23
1999	147,258	2.33
2000	154,011	2.44
2001	163,739	2.59
2002	165,456	2.62
2003	169,133	2.68
2004	165,917	2.63
2005	145,339	2.30
2006	181,577	2.87
2007	169,591	2.68
2008	173,146	2.74
2009	181,248	2.87
2010	232,173	3.68
2011	237,681	3.76
2012	266,866	4.22
2013	292,692	4.63
2014	315,886	5.00
2015	316,119	5.00
2016	324,649	5.14
2017	336,756	5.33
2018	260,712	4.13
2019	190,378	3.01
2020	92,260	1.46
Total	6,316,628	100.00

Table 5: Models to Match Patent Assignee and BR Establishment

Model	Assignee Name	State	City	Sequence
A1	Strict Name	Strict State	Strict City	1
A2	Strict Name	Strict State	Fuzzy City	2
A3	Fuzzy Name	Strict State	Strict City	3
A4	Fuzzy Name	Strict State	Fuzzy City	4
B1	Strict Name	Missing State	Strict City	5
B2	Strict Name	Missing State	Fuzzy City	6
B3	Fuzzy Name	Missing State	Strict City	7
B4	Fuzzy Name	Missing State	Fuzzy City	8
C1	Strict Name	Strict State	Missing City	9
C2	Fuzzy Name	Strict State	Missing City	10
D1	Strict Name	Strict State	Different City	11
D2	Fuzzy Name	Strict State	Different City	12
E1	Strict Name	Missing State	Missing City	13
E2	Fuzzy Name	Missing State	Missing City	14
F1	Strict Name	Different States	Same City (Strict or Fuzzy)	15
F2	Strict Name	Different States	Missing City	16
F3	Strict Name	Missing State	Different City	17
F4	Strict Name	Different States	Different City	18
G1	Fuzzy Name	Different States	Same City (Strict or Fuzzy)	19
G2	Fuzzy Name	Different States	Missing City	20
G3	Fuzzy Name	Missing State	Different City	21
G4	Fuzzy Name	Different States	Different City	22

Notes: The total number of all possible combinations by using strict and fuzzy name with state and city as blocking variables is 24 =2*3*4). That is, Strict or Fuzzy name * (Strict, Missing, or Different States) * (Strict, Fuzzy, Missing, or Different City). We have 22 in this table since "Same City" is identified as either a strict name matching or a fuzzy name matching.

Table 6: Preference Ordering of the Patent-level Match

Year Window	Sequence
appyear	1
gyear	2
appyear-1	3
gyear-1	4
appyear-2	5
gyear-2	6
appyear-3	7
gyear-3	8
appyear+1	9
gyear+1	10
appyear+2	11
gyear+2	12
appyear+3	13
gyear+3	14

Notes: We use three year gaps from a given reference year if there is no match found at the focal reference year. “appyear" or “gyear" refers to matches with application or grant year identical to the reference year; for each $k \in \{1, 2, 3\}$, “appyear-k" or “gyear-k" refers to the appyear or gyear being k years preceding the reference year; and “appyear+k" or “gyear-k" refers the appyear or gyear being k years after the reference year. The sequence number shows our priority.

Table 7: Models to Evaluate the Stem Name Matches

Model	State	City	Score
AA1	Strict State	Strict City	11
AA2	Strict State	Fuzzy City	10
BB1	Missing State	Strict City	9
BB2	Missing State	Fuzzy City	8
CC	Strict State	Missing City	7
DD	Strict State	Different City	6
EE	Missing State	Missing City	5
FF1	Different States	Same City (Strict or Fuzzy)	4
FF2	Different States	Missing City	3
FF3	Missing State	Different City	2
FF4	Different States	Different City	1

Notes: In the stem name matching, model types are defined only by the fuzziness of state and city, which gives us the above list. The order of our preference remains the same as before for the standardized name matching.

Table 8: Match Rates by Aggregate Model Types (%)

Model	Patent Level	Assignee Level
STD D+	62	55.5
STEM D+	14.8	12.8
ISRA	7	2.9
STD D- or STEM D- (No ISRA)	4.5	8.9
Overall	88.2	80.1

Notes: “STD D+" refers to the matches by standardized name based on the model A-D; “STEM D+" stands for those by stem name based on the model AA-DD; “ISRA" shows the matches by the internet search-aided algorithm; and “STD D- or STEM D- (No ISRA)" means the remaining matches by the model E-G or EE-GG (not identified by the search-aided algorithm). The model definition follows the same as in Table 5.