OECD Science, Technology and Industry Working Papers
2017/07

Matching Crunchbase with patent data

**Gianluca Tarasconi,
Carlo Menon**

<https://dx.doi.org/10.1787/15f967fa-en>

OECD SCIENCE, TECHNOLOGY AND INDUSTRY WORKING PAPERS

The release of this working paper has been authorised by Andrew Wyckoff, OECD Director for Science, Technology and Innovation and by Pierre Poret, OECD Director for Financial and Enterprise Affairs Directorate.

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to Directorate for Science, Technology, and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

© OECD 2017

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for commercial use and translation rights should be submitted to rights@oecd.org.

ABSTRACT

This note describes a procedure to match companies and individuals listed in Crunchbase, a new database on innovative start-ups and companies, with patent applicants and inventors reported in PATSTAT, the worldwide intellectual property database maintained by the European Patent Office. Given that neither administrative nor other unique identifiers are available in either of the two databases, the matching is based on a “fuzzy” procedure that exploits the available overlapping information across the two databases. A general overview of the resulting database completes the note.

TABLE OF CONTENTS

OECD SCIENCE, TECHNOLOGY AND INDUSTRY WORKING PAPERS	2
ABSTRACT	3
MATCHING CRUNCHBASE WITH PATENT DATA.....	5
Introduction	5
Data sources.....	6
The Crunchbase database.....	6
The PATSTAT database	6
A review of recent attempts of matching PATSTAT to firm repositories.....	7
Matching Crunchbase companies with PATSTAT applicants	8
The algorithm architecture	8
Matching individuals listed in Crunchbase with PATSTAT inventors	11
The algorithm architecture	11
Patenting start-ups and “startupper”: descriptive evidence.....	13
Patenting start-up	13
Patenting “startupper”	17
Conclusion.....	20
.....	20
NOTES	20
REFERENCES.....	21

Figures

Figure 1. Patenting start-ups by technological category	14
Figure 3. Distribution across patent offices	16
Figure 4. Share of patenting startupper by country and technological category	18
Figure 5. Share of patenting startupper by job title and gender	19

MATCHING CRUNCHBASE WITH PATENT DATA

Gianluca Tarasconi and Carlo Menon

Introduction

This note describes a procedure to match Crunchbase, a new database on innovative start-ups and companies, with information on intellectual property (IP) contained in PATSTAT, the worldwide database on IP maintained by the European Patent Office (EPO). The match covers both companies and inventors.

Given that neither administrative nor other unique identifiers are available in either of the two databases, the matching is based on a “fuzzy” procedure that exploits the available overlapping information across the two databases: the company names, their location, the names of the people linked to them. The matching procedure needs to be carefully designed in order to maximise the number of correct matches, while at the same time minimising both “false positive” and “false negative” errors. This is not straightforward as the spelling of companies’ and people’s name is not always consistent across the two databases. Furthermore, in the PATSTAT database there is neither a unique internal identifier for either applicants or inventors, which have therefore to be adequately disambiguated before engaging in the matching exercise. This is particularly critical for inventors, where homonymy is very frequent (the so-called “John Smith” problem). The sheer size of the PATSTAT database (several million patents, distinct applicants, and inventor¹s) makes the procedure burdensome also from a computational point of view.

For this version of the matching, the priority has been given to precision - that is, minimising type I errors, or false positives. However the algorithm can be calibrated differently, e.g. with the intent to maximise recall (minimise type II errors, or false negatives), or to possibly obtain any Pareto-optimal combination of recall and precision (that is, to strike a balance between different types of errors; Pezzoni, Lissoni, and Tarasconi, 2014).

Almost 50 000 companies, out of the 447 000 listed in Crunchbase (excluding venture capital companies), are found to own one or more patents, for a total of around 12 million patents. Around 220 000 of those have been applied for by companies created after 2005. The share of patentees for US companies is 15%, but the share doubles for companies reporting at least one funding round.

Regarding individuals, out of the 578 000 professionals listed in Crunchbase who could be potential patent inventors, around 25 000 are found to have a correspondent in PATSTAT. These inventors account for 2.2 million patent applications.

Data sources

This Section contains a brief description of the two main data sources used in the matching. The interested reader can refer to Dalle, Denbesten, and Menon (2017) for a more comprehensive description of Crunchbase and of its potential for academic and policy research.

The Crunchbase database¹

Crunchbase is presented in its website as "Crunchbase is the destination for discovering industry trends, investments, and news about businesses, from startups to the Fortune 1000. Crunchbase provides a depth and breadth of knowledge that ensures its data is recognised as the primary source of business information by over 31 million users globally²".

In the version used for this note, downloaded in January 2017, the database contains information on more than 490 000 distinct entities located in 199 different countries; 447 000 are companies, while the remaining ones are VC investors, schools, or business groups. Of those 447 000 companies, 137 000 report a founding year later than 2010, and 210 000 later than 2005.

The data are sourced through two main channels: a large investor network and community contributors. More than 3 000 global investment firms submit monthly portfolio updates to Crunchbase, in exchange of free data access. In addition, around 500 000 executives, entrepreneurs, and investors contribute to update and revise Crunchbase company profile pages. This wealth of data is processed with artificial intelligence and machine learning algorithms in order to ensure accuracy and scan for anomalies. Additionally, algorithms continuously search the web and thousands of news publications for information to enrich profiles³. The database started being populated in May 2007, and for every company the date in which the related record was created is reported. The pace of new record creation was however rather limited until the beginning of 2013, when it stabilised at around 260 records per day, on average, with the exceptions of two dates in 2013 and 2014, respectively, when around 6 000 and 8 000 records were added, probably as the consequence of the acquisitions of additional sources. Given that Crunchbase is quickly becoming a primary data source for investors, it is plausible to assume that dynamic and ambitious entrepreneurs have a strong incentive to register on the website and to keep their information updated.

The coverage appears particularly comprehensive for start-ups located in the United States and operating in the e-commerce, advertising, and mobile app sectors. However, the share of companies with at least one registered VC deal is much higher in sectors like biotech and health care. The majority of funding, however, is concentrated in the biotechnology sector.

Crunchbase also contains around 580 000 records on people who are connected to at least one company listed in the database. The following variables are reported: the full name, location (city and region), gender, job title, and the dates on which the record was created and updated, respectively. Most people are classified as founder, co-founder, and CEO. The distribution across categories mirrors that of companies, but with some differences in the relative ranking.

The PATSTAT database

PATSTAT is the short name for "EPO worldwide PATent STATistical Database", a single database covering 100 million patents from 90 patent authorities⁴, which has been developed by European patent Office (EPO) in cooperation with WIPO, OECD and Eurostat.

PATSTAT is distributed twice a year in a set of 30 tables that create a relational database, allowing advanced, large scale statistical analysis of patent data. Information is extracted from EPO's master bibliographic database DOCDB, the EPO's worldwide patent information resource.

DOCDB is an “examiner centred” dataset: data are collected mainly to be helpful to identify prior art for the examination process; thus information such as technological classes and citations is normalised and more complete compared to the fields which are marginal for that purpose (like applicants' or inventors' names or addresses). For such reasons data extracted from PATSTAT have been extensively preprocessed, as described in following paragraphs, in order to clean, normalise applicants names and, to a certain extent, to refill applicants addresses.

PATSTAT, as a relational database, is built based on the patent application unique identifier as the main surrogate key, to which the full sets of bibliographic variables are linked; the most important are:

- Priority date, application and publication number and date
- Title and abstract of the application
- Status of application (granted, pending)
- International Patent technological Classification (IPC) codes
- Applicant's name and address
- Inventors' names and addresses
- References (citations) to prior-art patents and to non-patent literature
- Patent Family⁵

A review of recent attempts of matching PATSTAT to firm repositories

Previous attempts to link patent data to other dataset mainly focus on matching with companies listed in the Bureau van Dijk databases (ORBIS, AMADEUS) or Compustat. Hall, Jaffe, and Trajtemberg (2001) in their seminal work matched a dataset of patents applied at USPTO from 1963 to 1999, combining name harmonisation with manual matching, in order to minimise both false positives and false negatives. This type of approach, as well as the routines and code produced, proves to be particularly effective with US firms, in particular with big corporations.

An attempt to match PATSTAT data with the commercial database ORBIS (maintained by the Bureau Van Dijk) is described in Thoma et al. (2010). The matching is based on automatic routines and algorithms to harmonise, disambiguate, and match applicants with firms in ORBIS. Particular effort was put in removing noise arising from typing mistakes, and in standardizing name conventions. The exact matching cases are thus complemented with approximate matching, where information on the location of applicants and firms is used to identify false matches, and scores are computed both in terms of string similarity functions and token distance. Due to the sheer number of matches, no global manual check was performed. Ultimately 131,065 companies included in ORBIS were identified as EPO applicants, corresponding to about one million EPO applications in 1979-2008. Part of the cleaning and matching methodology used in that endeavour has been replicated for this matching exercise.⁶

A similar exercise is described in Squicciarini and Dernis (2013), and resulted in the OECD ORBIS-PATSTAT matched database. In a first step, the names of firms included in the patent and the ORBIS databases are separately harmonised using country-specific dictionaries, which prove to be particularly useful especially for Asiatic countries. A series of string-matching algorithms, mainly token-based and string-metric-based, then compare the harmonised names from the two datasets and provides a matching accuracy score for each pair. The result of the matching is then checked manually. As in many cases a patenting firm may correspond to several entries in the ORBIS database, a series of disambiguation rules has been implemented, e.g. by comparing company addresses, and by giving priority to companies with the largest turnover figures.

Another important contribution was given from a joint study from European patent office and Office for Harmonisation in the Internal Market (OHIM) (EPO-OHIM, 2013). In this case a standard preprocessing and cleaning was applied, especially aimed at cleaning and filtering legal forms in European states. This methodology however is not replicable in this case since the majority of the data examined are from non-European countries and patent authorities.

A totally different approach is described in Balsmeier et al. (2016), where Machine Learning methodologies have been used for disambiguating and standardizing applicants names. In a first step, a dictionary was built using the NBER matching; in a second step, new assignees' names are first standardised and queried for in the lookup dictionary constructed from the NBER data; in case no correspondence is found, a K Nearest Neighbours classifier is used. The method proved to be very effective, however it cannot be replicated in the case under analysis here as the NBER dictionary is not appropriate for the population of firms in Crunchbase, and no 'training set' is available to set up a machine learning procedure.

Matching Crunchbase companies with PATSTAT applicants

The matching of the Crunchbase company list with PATSTAT presents a few additional challenges as compared to previous similar attempts, as highlighted in the previous Section. More importantly, the novelty of the matching exercise depends also on the joint availability of people and company names. While there are documented examples of attempts of matching scientists' or researchers' names from various databases with names of inventors listed in PATSTAT, this is the first systematic and large-scale attempt at matching the two kind of records jointly, to the best of the authors' knowledge.

The algorithm architecture

In order to enrich Crunchbase data with patent information, the database has been matched with patent data taken from PATSTAT autumn 2016 edition. PATSTAT is a database delivered from EPO containing bibliographical data relating to more than 90 million patent documents from leading industrialised and developing countries. This type of match presents many challenges, notably arising from the way data on applicants are stored in PATSTAT. The first issue is the lack of harmonised names, which means that the same entity may have several separate database entries (due also to the many different spellings of a single organisation or name changes over time). Another issue is the lack of comprehensive information about applicants (only address information is available, and still this piece of information is not standardised and often partial or missing). In addition, the distribution of the number of patents per assignee is skewed; a very small number of applicants hold thousands of patents, while the large majority of assignees are small entities that hold less than five patents.

Selection of the relevant samples

Only applications filed for at the European, US, Korean, Chinese, and Japanese patent offices (IP5: EPO, USPTO, KIPO, SIPO and JPO) after year 2000 have been taken into account. The restriction to IP5 applications is a common practice in the literature, as those are universally recognised to be the most important patent offices worldwide. Applicants that are classified as individuals, rather than companies, were dropped. For Crunchbase, only companies listed as venture capital firms (i.e., investors) were dropped. The basic entity of the match is the couple company name and country code. Even if full address information is available in Crunchbase, it could lead to a potential bias since in PATSTAT the coverage of this information is satisfactory only for EPO applications, and totally missing for JPO or SIPO. On the other hand, given the high percentage of start-ups contained in Crunchbase, changes of address are very likely to take place in the first years of the life of a company, which leads to a high risk of false negatives.

Refilling country codes in Crunchbase and PATASTAT

In the selected sample, the country code is missing for about a 30% of patent applicants in PATSTAT and a 25% of companies from Crunchbase. For PATSTAT, the refilling strategy builds on the methodology developed by de Rassenfosse (2013) based on the following steps:

- Find a homonym in the same patent family.
- If more than one country code is found, the country of the applicant with the higher number of patents (over the full PATSTAT database) is assigned.
- If no homonym is found, if the applicant belongs to a patent family of only one patent (singleton), the nationality of patent office is given (this case helps disentangle in particular cases of SIPO and JPO only applicants).

This algorithm solved the majority of cases, leaving only about 2500 names (less than 1%) with no country code. For Crunchbase, the refilling of the missing country codes is done using i) the modal country code of the people reported to work for the company, and ii) the telephone country code, whenever available and unambiguous.

Harmonisation of applicant and company names

As previously mentioned, one of the major challenges of the matching is reconciling different names to the same legal entity. For applicants' names, the work described here takes stock of the work of Magerman et al. (2013) by using the standardised names from EEE-PPAT database, now included in PATSTAT itself. In addition, the names extracted are further processed by removing the remaining "noise" and most of the legal designations (e.g. LLC, CORP, INC) from the entity names (as it is done in e.g. Vezzani et al., 2014). Data from both sources have passed this process phase in order to ensure that the results are comparable.

String matching

In the name match phase, four criteria are combined, listed below in order of increasingly match accuracy:

1. Perfect match: where names, removing legal designation, are exactly the same.
2. Alphanumeric match: where the names, keeping only [A-Z] and [0-9] are the same (e.g.: I.B.M. = IBM = I B M).
3. Jaro-Winkler distance: names are broken into tokens and the similarity score is computed by the number of tokens in common, weighted on the inverse of frequency. The higher the Jaro-Winkler distance for two strings, the more similar the strings. Only results above a threshold value have been considered valid matches.
4. Levensthein distance: this measure is an edit distance counting the number of changes needed to transform the former name into the latter.

The latest criterion eventually proves to add a high number of false positives because in most cases (e.g. acronyms) by changing one letter the algorithm would match two very different entities; thus the matching is kept only if can be validated with inventors' names, as explained below. Each match receives a score that is proportional to both the accuracy of the type of matching and, for the third and fourth criteria, the distance between the two names.

As previously stated, aside from names, country codes have also been taken into account as main criteria for match, because the same applicant name in different nations may refer to distinct entities (e.g. "Ministry of Health").

Additional filtering using information on people

Subsequently, the list of people affiliated to all companies matched in the previous step is compared with the list of patent inventors available in PATSTAT. The possible similarity of people names is particularly useful to disambiguate those cases in which the same patent applicant in PATSTAT is matched to different companies in Crunchbase, or whenever the company name matching does not have a high precision score. Overall, around 7% of matches, corresponding to 11% of Crunchbase companies, are validated using information on inventors.

Benchmarking the matching on the Bureau Van Dijk ORBIS database

Whenever possible, patent matching exercises are typically benchmarked using a subsample of the matched sample for which other direct sources of information without error are available, generally from surveys or similar sources. These benchmarking exercises are generally based on two indices: the precision rate, i.e., the proportion of true links from all predicted links, and recall rate, i.e., the proportion of true links that are correctly predicted by the algorithm (see e.g. Pezzoni, Lissoni, and Tarasconi, 2014).

In this case, a sample reporting the "true" matches is not available. However, the resulted matched database can be benchmarked with the ORBIS database, which also reports information on patents. The comparison is based on a small overlapping sample of 7.569 companies that matches exactly and unambiguously by the company name and country code. The precision and recall rates calculated on the overlapping patent sample are both above 90%. Considering that the matching

executed by the Bureau Van Dijk is also “fuzzy”, being based on company name similarity, and it may thus also contain a reasonable error margin, these rates are more plausible (and reassuring) than 100% precision and recall rates. In addition, the procedure described in this note is able to find another 2.913 patents, corresponding to 8% of the overlapping patent sample.

Matching individuals listed in Crunchbase with PATSTAT inventors

The second part of the matching exercise consists in finding a correspondence between individuals listed in Crunchbase and patent inventors listed in PATSTAT. Given the sheer number of distinct inventor names and the higher incidence of homonymy with people than with companies, the people matching exercise requires a more complex design than the company one. In particular, it is necessary to effectively filter the sample of potential matches in the PATSTAT database for each individual listed in Crunchbase. This Section describes the procedure in detail.

The algorithm architecture

The algorithm is structured around the following steps: selection of the relevant people sample in Crunchbase; refilling of missing country codes in Crunchbase; harmonisation of inventors’ names in PATSTAT; name fuzzy matching; and additional filtering using complementary information, namely the fields of technological specialisation and current and previous employers’ names.

The procedure applied is mainly based on Pezzoni et al. (2014). The main difference is that, while their procedure aims to disambiguate names within PATSTAT itself, this algorithm matches two different datasets. The complementary information that has been used to filter the name matches are: geographic areas (Region, county, state); patent applicants/employers; network information (coinventors in common; networks with three degrees of distance); technological area (patents IPCs, Crunchbase technological categories).

Since inventors in PATSTAT are not classified with a unique individual identifier, the algorithm first matches and filters PATSTAT inventors only, in order to disambiguate different entries that refer to the same individual. Once this is done, the algorithm matches the PATSTAT disambiguated inventors’ list with individuals listed in Crunchbase, who are already disambiguated in the source database. This is important to avoid false negatives, e.g. in the case in which inventor A and B listed in PATSTAT are the same individual, but only B matches the Crunchbase entity C, because the additional filtering information (e.g. previous applicant/employer and technological category) that matches with those reported in Crunchbase for C are available only for B, and not for A.

Selection of the relevant samples

The first step consists in filtering out individuals in Crunchbase with a job title that is not generally associated with patenting. These include occupations in finance, marketing, and sales. PATSTAT inventors have been selected among those who have an application filed at IP5 offices after 1978.

Harmonisation of names

Names harmonisation has followed standard textmining criteria, removing punctuations, double-spaces, putting all names in uppercase and converting all characters who are part of extended charset (accented letter, umlauts etc.) to the basic Latin corresponding chars. For PATSTAT names, where possible also company names (often included as c/o in inventors names) and a long list of academic and study titles, both international and country-specific, (e.g.: Dr., PhD, Ing.) have been removed.

Name clustering based on string matching

The name matching and clustering phase create all the possible couples of Crunchbase individuals and PATSTAT inventors that are similar enough to be considered as possible homonyms, leaving to the following filtering phase the task to drop matches between two entities that do not refer to the same individual. Due to the higher international mobility of inventors, compared to companies, the match has been performed by name, ignoring other geographic information. This is motivated by the fact that, for inventors, address levels below the country code (street, city, zip, etc.) are missing in PATSTAT for most patent offices but EPO. Country code is missing as well for a high share of inventors data, and refilling may imply loss of possible matches for individuals who moved their residence during their career.

Crunchbase data have already a good name and surname parsing, where PATSTAT may have the two of them inverted, leading to a choice of an algorithm based on tokens similarity. However, misspelling or alternative forms for names (e.g., Rob vs. Robert), and third names listed, omitted, or added just with the first letter, would all lead to drop a high percentage of valid matches by using token similarity alone. Therefore, token similarity has been extended by using 2grams similarity, an edit distance measure that counts the number of two characters pieces that two strings have in common, divided by the number of bigrams of the two strings. More formally, it can be described as the vector distance between two strings of different lengths, normalised by the total length of the strings.

For example, if one were to consider a very common misspelling like "adminstration" vs "administration", by breaking the words in two letter pieces one obtains

ad dm mi in ns st tr ra at ti io on

ad dm mi in ni is st tr ra at ti io on

which leaves with 11*2 tokens in common out of 25, corresponding to a 0,88 similarity.

Disambiguation of PATSTAT inventors

The clustering of PATSTAT inventors who show some degree of similarity with at least one Crunchbase individual creates 300 million couples, corresponding to 9.8 million PATSTAT person IDs. The filtering algorithm follows a customised version of the “balanced precision recall” procedure described in Pezzoni et al. (2014). Namely, the two entities must have in common at least three criteria, out of the following five: at least one applicant in common; at least one common IPC4 tag; having one applicant with less than 50 inventors; at least one coinventor in common; and being at maximum three degrees of distance in patenting⁷.

CrunchBase-PATSTAT match

The disambiguation of the PATSTAT inventors produces 14.9 million possible matches between the sample of Crunchbase staff and the sample of disambiguated PATSTAT inventors. These matches are filtered based on the first three criteria described above, i.e., at least one applicant in common; at least one common IPC4 tag; and having one applicant with less than 50 inventors. The last two criteria are instead dropped because coinvention network information is not available in Crunchbase. As with the previous step, a match is validated if at least two of the above criteria are satisfied.

Patenting start-ups and “startupper”: descriptive evidence

This Section presents a preliminary snapshot of patenting start-ups, in order to give a flavour of the potentialities of the combined CrunchBase-PATSTAT database for economic research. More refined descriptive and econometric analysis is left for future work. In particular, forthcoming work has to take into account the possible selection bias originating from the varying coverage of start-ups in Crunchbase. It is therefore important to keep in mind that the patterns discernible from these graphs may depend also from sample composition issues. To easy visibility, in the graphs only the 20 largest countries or technological categories in terms of start-ups (Figure 1) or patent counts (Figures 2 to 4) are reported. Furthermore, traditional caveats for the analysis of patent data also need to be taken into account (see e.g. OECD, 2009).

Patenting start-up

Figure 1, Panel A, shows the number of patents by technological category, measured both as simple counts and weighted by the inverse of the family size. Start-ups in the biotechnology sector are responsible for the largest number of patents by a large margin, however this is partly explained by the larger family size of their patents; once this is taken into account, the difference with other categories is less striking. The Panel B of Figure 2 shows that in some technological categories, like e.g. e-commerce and smartphone applications, the amount of VC funding is disproportionately going to patentees. This finding is confirmed by unreported regression analysis, in which the positive relationship between being a VC recipients and owning at least one patent is significant even controlling for country and technological category fixed effects.

Figure 2 reports the average - by country and technological category - of two indicators of patent quality attributes (Squicciarini, Dernis, and Criscuolo, 2013). The first indicator is the share of backward citations to non-patent literature and is typically interpreted as an indicator of closeness to science. The second indicator is a proxy for the patent radicalness and is calculated as the number of different International Patent Classification (IPC) classes cited that are different from the one the citing patent belongs to (i.e., radical patents are those that bring insights from different technologies). The two indicators are available only for patents filed at the USTPO and EPO, and therefore the sample is restricted accordingly. In the graph in Panel A, Denmark appears to be the country where start-up patents are closer to science, while Chinese Taipei is the economy in which start-up patents are more radical. Across technological categories (Panel B), biopharma start-up patents are the closest to science, while start-ups operating in b2b and chemicals appear to develop the most radical patents, in average.

Finally, Figure 3 shows the shares of patent applications by patent offices, across countries and technological categories, respectively. The graphs show a clear “home effect” with start-ups filing for patents more frequently at the national (or regional, in the case of EPO) patent office. For instance, almost 80% of patents owned by US start-ups are filed at the USTPO; the latter patent office is also a popular choice of Canadian and Israeli start-ups. Conversely, European start-ups tend to file for more frequently at the EPO.

Figure 1. Patenting start-ups by technological category

Panel A: number of patents

Panel B: funding to patentees vs. non-patentees

Note: the sample is limited to companies less than 15 years old.
 Source: <http://www.crunchbase.com> and PATSTAT

Figure 2. Closeness to science and radicalness

Panel A: by country

Panel B: by technological category

Note: the sample is limited to companies less than 15 years old.
 Source: <http://www.crunchbase.com>, PATSTAT, and Squicciarini, Dernis, and Criscuolo (2013).

Figure 3. Distribution across patent offices

Panel A: by country

Panel B: by technological category

Note: the sample is limited to companies less than 15 years old.

Source: <http://www.crunchbase.com> and PATSTAT

Patenting “startuppers”

While the previous Section looked at patenting activity across firms, this Section presents a quick description of patenting activity across the individuals listed in Crunchbase as working for a start-up (named with the neologism “startuppers” in this report). The same caveats of the previous Section hold: this preliminary work is only intended to give a flavour of the database potentialities.

Figure 4 reports the share of inventors - i.e., the share of individual for whom it was possible to find a correspondent in PATSTAT - by country (Panel A) and technological category (Panel B). Israel appears to be the country with the highest share of start-uppers who are also patent inventors, with a share of around 8% (Panel A). Looking at technological categories, the graph in Panel B shows that more than one out of five start-uppers working in biotechnology are patent inventors.

Figure 5 visualises the share of patenting start-uppers by job title (Panel A) and gender. Interestingly, those start-uppers who report their job title to be founder, president, co-founder, chief executive officer (CEO), and chief technology officer (CTO), appears to be significantly more likely to be inventors, with a share higher than 10%. Panel B of the same figure instead shows that the share of inventors is higher for male start-uppers than for female ones. It would be interesting to investigate the issue further, e.g. by assessing whether the different gender ratio across job titles, countries, and technologies can at least partially explain that.

Figure 4. Share of patenting startupper by country and technological category

Panel A: by country

Panel B: by technological category

Note: the sample is limited to companies less than 15 years old.

Source: <http://www.crunchbase.com> and PATSTAT

Figure 5. Share of patenting startupper by job title and gender

Panel A: by job title

Panel B: by gender

Note: the sample is limited to companies less than 15 years old.

Source: <http://www.crunchbase.com> and PATSTAT

Conclusion

This note describes a procedure to match companies and individuals listed in Crunchbase, a new database covering innovative start-ups and companies around the world, with the EPO World Patent Database. The procedure allows obtaining a database of around 50 000 patenting companies, owning a total of 12 million patents, and 25 000 individual inventors, who filed for 2.2 million patents.

The resulting database can be used in a number of different domains. These include the analysis of the role of IP assets in securing venture capital; the characterisation of the IP portfolio of high-growth patenting start-ups, of start-ups developing radical or breakthrough innovations, and of inclusive start-ups; the analysis of the linkages between patenting innovative start-ups and public research, exploiting patent citations to other patents, as well as to non-patent literature (NPL).

Given the international dimension of both the Crunchbase and PATSTAT database on the one hand, and the richness and granularity of the available micro-data for both databases, on the other hand, it is easy to recognise the potential of the matched database for meaningful comparisons across countries, technologies, and types of start-ups.

NOTES

¹ This draws heavily on Dalle, Denbesten, and Menon (2017).

² <http://about.crunchbase.com/> visited on September 11th, 2017

³ <https://about.crunchbase.com/products/the-crunchbase-difference/> visited on September 11th, 2017

⁴ Information on access, documentation and coverage can be found at <https://www.epo.org/searching-for-patents/business/patstat.html>

⁵ Details and different types of families are described at: <http://www.epo.org/searching/essentials/patent-families.html>⁶ In the matching exercise described in this paper the geographic filtering approach is not implemented, because of both the lack of this information in Patstat for patents applied outside EPO, and the fact that in Crunchbase only the last address of the headquarter is listed.

⁷ Any pair of inventors I and J are said to stand at three degrees of separation when at least one of I's coinventor and one of J's coinventor have collaborated on the same patent.

REFERENCES

- Balsmeier, B., G.C. Li, T. Chesebro, G. Zang, G. Fierro, K. Johnson, A. Kaulagi, S. Lück, D. O'Reagan, B. Yeh,. And L. Fleming (Forthcoming). "Machine learning and natural language processing on the patent corpus: data, tools, and new measures". *Journal of the European Mathematical Society*, Forthcoming.
- Dalle, J.M., M. Denbesten, and C. Menon (2017), "Using Crunchbase for academic and managerial research". *OECD Science, Technology, and Industry Working paper*, Forthcoming.
- OHIM, EPO (2013), Impact of intellectual property rights intensive industries: contribution to economic performance and employment in the European Union. Industry-Level Analysis Report.
- OECD (2009), "The OECD patent statistics manual." *OECD Publications*, Paris.
- Pezzoni, M., F. Lissoni and G. Tarasconi (2014), "How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation". *Scientometrics*, 101(1), 477-504.
- Squicciarini, M., H. Dernis and C. Criscuolo (2013), "Measuring patent quality: Indicators of technological and economic value". *OECD Science, Technology and Industry Working Papers*, 2013(3), OECD Publishing, Paris. DOI: <http://dx.doi.org/10.1787/5k4522wkw1r8-en>
- Squicciarini, M. and H. Dernis (2013), "A Cross-Country Characterisation of the Patenting Behaviour of Firms based on Matched Firm and Patent Data", *OECD Science, Technology and Industry Working Papers*, No. 2013/05, OECD Publishing, Paris. DOI: <http://dx.doi.org/10.1787/5k40gxd4vh41-en>
- Hall, B. H., A. B. Jaffe and M. Trajtenberg (2001), "The NBER patent citation data file: Lessons, insights and methodological tools". *National Bureau of Economic Research Working Paper* no. 8498.
- Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B.H. and Harhoff, D. (2010), "Harmonizing and combining large datasets-An application to firm-level patent and accounting data". *National Bureau of Economic Research Working Paper* no. 15851