Research Policy 38 (2009) 1617–1627

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

How to play the "Names Game": Patent retrieval comparing different heuristics

Julio Raffo[a,b,*], Stéphane Lhuillery[a]

[a] CEMI-MTEI, Collége du Management de la Technologie, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[b] CEPN, Université Paris Nord, France

ARTICLE INFO

Article history:
Received 3 June 2008
Received in revised form 19 July 2009
Accepted 10 August 2009
Available online 23 September 2009

JEL classification:
C63
C81
C88
O34

Keywords:
Patents
Inventors
Names matching algorithms
Indicators
PATSTAT

ABSTRACT

Patent statistics represent a critical tool for scholars, statisticians and policy makers interested in innovation and intellectual property rights. Many analyses are based on heterogeneous methods delineating the inventors' or firms' patent portfolios without questioning the quality of the method employed. We assess different heuristics in order to provide a robust solution to automatically retrieve inventors in large patent datasets (PATSTAT). The solution we propose reduces the usual errors by 50% and casts doubts on the reliability of statistical indicators and micro-econometric results based on common matching procedures. Guidelines for researchers, TTOs, firms, venture capitalists and policy makers likely to implement a names game or to comment on results based on a names game are also provided.

1. Introduction

The demand for scientific and technology (S&T) indicators has grown over the past 20 years. The initial efforts were concentrated on the S&T activities of firms and led statistical offices and institutes to launch and harmonize R&D and innovation surveys (see Freeman and Soete, 2009). Some institutions (e.g. NSF, AUTM) did this in order to collect information on public research organizations but, beyond some efforts undertaken on R&D activities (see OECD, 2002), no systematic or coordinated effort has been made at an international level. Therefore, despite the importance of PROs in national systems of innovation, little is known about their activities compared with firms. The persistent lack of information is problematic as policy makers and PRO managers increasingly wish to assess and monitor the productivity of public research organizations in a systematic way. The demand for indicators is particularly important in Europe where policy makers are eager to know how much the EU is lagging behind US universities in terms of publications, inventions and technology transfer. The need for additional and proper data is not limited to policy makers or PROs. It is also of interest to firms and venture capitalists looking for technological opportunities and trying to circumvent their competitors' technological threats (e.g. through blocking patents).

This rising demand for indicators associated with improved access to IPR data, has rejuvenated the interest in patent statistics produced by scholars, especially as these latter are now able to match large patent datasets with any list of university employee or academic author names (Trajtenberg et al., 2006; Hoisl, 2006; Thursby et al., 2009). This "names game" – as named by Trajtenberg et al. (2006) – is a first and mandatory step in the accurate assessment of agents' or organizations' patent portfolios. It thus determines the subsequent analyses of inventivity approximated by patent count (e.g. Balconi et al., 2004), and the identification of technological networks or technological profiles (e.g. Cantner and Graf, 2006).

The methodological problems associated with these matching procedures have been however largely overlooked: first, scholars still don't know what the best procedures to match different datasets are; second, scholars and policy makers are unaware of the level of influence of the chosen solutions on the final results. In other words, there is a potential problem of reliability which has not been not alleviated by the recent efforts of transparency made by authors in their matching methodology (e.g. Thursby et al., 2009).

* Corresponding author at: EPFL, CDM-MTEI-CEMI, Odyssea-Station 5, 1015 Lausanne, Switzerland.
E-mail address: julio.raffo@epfl.ch (J. Raffo).

0048-7333/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.respol.2009.08.001

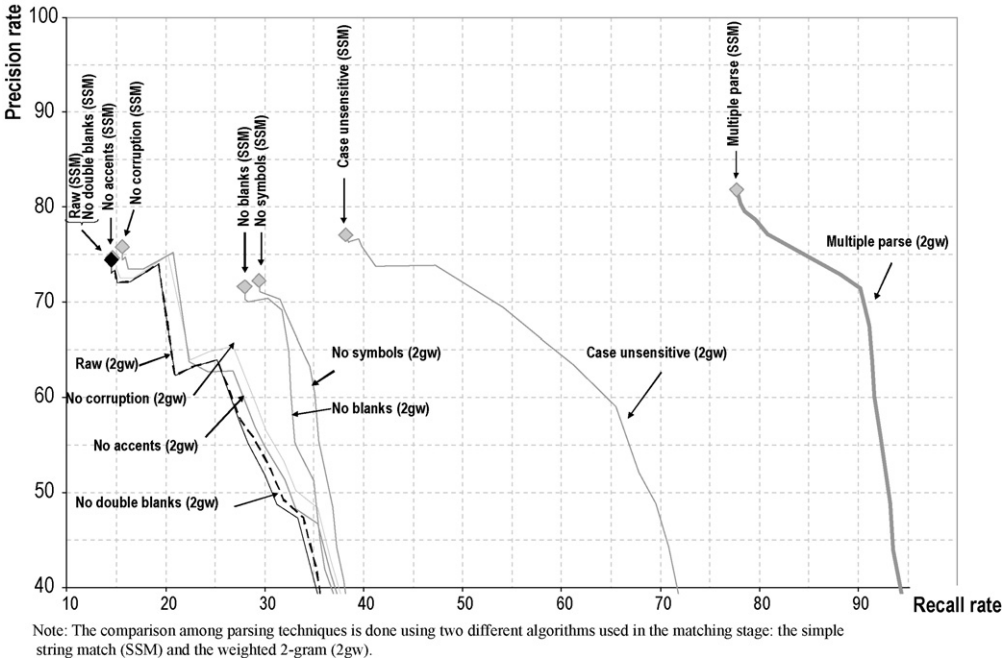


Fig. 1. The impact of different parsing techniques on the precision-recall frontier. *Note*: The comparison among parsing techniques is done using two different algorithms used in the matching stage: the simple string match (SSM) and the weighted 2-gram (2gw).

One main reason for this oversight is that the analysis of a "names game" is a very complex task. Any name matching procedure can be conceptualized as having three sequential stages: a first "parsing" stage aimed at cleaning up any noise such as different cases, corrupted characters, double spaces, etc. in patent databases and researchers' or firm names' lists. The "matching stage" consists in applying a matching algorithm to obtain a list of potential matched pairs. In the final "filtering stage", complementary information is used in order to disambiguate true matches from false ones. There are two major difficulties with this three-stage names game: the first is the choice of steps or procedures to apply within each stage. The second concerns the procedure's sequencing, as the choice made for each step determines firstly the problem to solve and consequently the performances of procedures applied at the following stages. The identification of the best procedures sequence is therefore not straightforward in a problem where there is no silver bullet solution.

Even though there are no straightforward solutions, the present paper proposes to evaluate several possible heuristics, comparing possible alternatives inside the three stages and the interactions between them. Using a large patent dataset (the European Patent Office (EPO) Worldwide Patent Statistical Database called "PATSTAT" hereafter), we identified patents for a set of inventors from the Ecole Polytechnique Fédérale de Lausanne (EPFL hereafter). The EPFL Technology Transfer Office (TTO) provided a list of 349 inventors listed in EPFL patents for the period 1995–2005. The TTO information on these inventors was complemented with the data from the EPFL human resources office. The combination of the two sources enabled us to build a small but precise benchmark set of EPFL inventors and their patents (1830 pairs). This set was then used in order to assess the performances of different heuristics and their impact on descriptive statistics and micro-econometric studies.

The paper is organized as follows. Section 2 introduces the three different stages and investigates the performances of different algorithms implemented for each stage. Section 3 presents the best heuristic we found and assesses its robustness. We propose in Section 4 presents the implications of our results for micro-econometric results and statistical indicators, used by policy makers and scholars. A final section discusses our conclusions.

2. A three-stage game

2.1. The parsing stage

The parsing stage is a data preparation strategy aimed at reducing the noise in the name field (e.g. address, institution, title) without removing information which sometimes can be useful in subsequent stages. For example, a middle name is additional information in patent data sets (about 21% in our benchmark set) which provides a lot of useful information when disentangling homonyms but which also creates serious problems when the matching algorithms applied after the parsing stage are not able to identify the similarity between for example "Luis Egidio Miotti" and "Luis Miotti". Other problems such as different case (found in 72% of names on the EPFL TTO list), symbols (18%), accentuated characters (15%) and double spacing (14%) are frequent and easier to deal with through systematic parsing, when applied to the two different datasets to be matched.

Fig. 1 highlights the impact of the parsing stage on the final matching results by comparing data after seven different parsing techniques and the original data with no parsing (labeled as "Raw" in Fig. 1). The precision–recall points and curves[1] in Fig. 1 raise three crucial aspects to consider when applying parsing techniques.

Firstly, such techniques impact differently on the matching results. More precisely, Fig. 1 illustrates that when the techniques are applied one by one, the highest impact comes from the transfor-

[1] When a Type I error (or false negative) occurs, it decreases the Recall rate whereas a Type II error (or false positive) decreases the Precision rate. Recall rate is defined as CR/(CR + CM) where, CR is Correct Recall, CM is Correct Missing (Error type I or false negative) and Precision Rate as CR/(CR + IR) where, IR is Incorrect Recall (Errors type II or false positive). It is usually considered in the literature that the higher both precision and recall are, the better the matching is.

mation to the same case (all lower or all upper case). This suggests that there is strong case heterogeneity in the way names are written in patent documents and that the gains achieved in standardizing it overrule any interest in keeping the information contained in a case sensitive string, such as the initials of first and middle names. This parsing procedure prevails even though the cleaning-up of symbols – such as dots, commas, hyphens or apostrophes – also results in a large increase in recall rates, with a minor decrease in precision. Accordingly, if some precision loss is tolerable, considerable recall improvements are obtained from removing all blank spaces – a parsing technique that has already been used for matching firms (Magerman et al., 2006) – although this technique is incompatible with some matching algorithms like the Token based ones. Furthermore, removing accents, double blanks or other corrupted characters are parsing tasks which, if applied separately, do not greatly improve precision or recall.

Secondly, although the results from the different parsing techniques are similar in the way they impact, their distinct impacts are not equal as they depend on the subsequent matching algorithm employed, Evidence also suggests the existence of synergies when applying several parsing strategies all together. As can be seen in Fig. 1, the rightmost curve reflects the results of applying all parsing techniques (except "No blanks") and the gains in terms of precision and recall are much greater than those provided by the simple addition of each parsing technique's marginal gain. The combination of parsing techniques we propose here results in improved precision (+7%) and recall rates (+64%) suggesting that scholars working on patent matching have good reason to insist on the importance of the parsing stage in their papers (Magerman et al., 2006).

2.2. The matching stage

With respect to the matching stage, the available literature using patent data favours the "simple string match" algorithm (see however Kim et al., 2005; Trajtenberg et al., 2006; Thoma and Torrisi, 2007). Still, as Appendix A demonstrates, many matching algorithms exist and can be applied to inventors' names or even firms' names. Among them, we selected the following different algorithms: simple string match, Soundex, Metaphone, Edit distance, 2-Gram, 3-Gram and Token algorithms[2] which we applied to our multi-parsed benchmark set in order to assess their relative performances.

Our main results are represented in Fig. 2, where the Recall–Precision results for simple string match, Soundex and Metaphone are represented by single points while Edit distance, N-grams and Token are represented by decreasing curves. These curves characterize the Recall–Precision trade-off when changing the value of the algorithm similarity threshold. The closer the threshold is set to one, the higher the Precision rate will be, while the lower it is set, the higher the Recall rate will be.

Given its restrictive nature, the simple string match (SSM) is believed to be the best algorithm in terms of precision rate (82%)

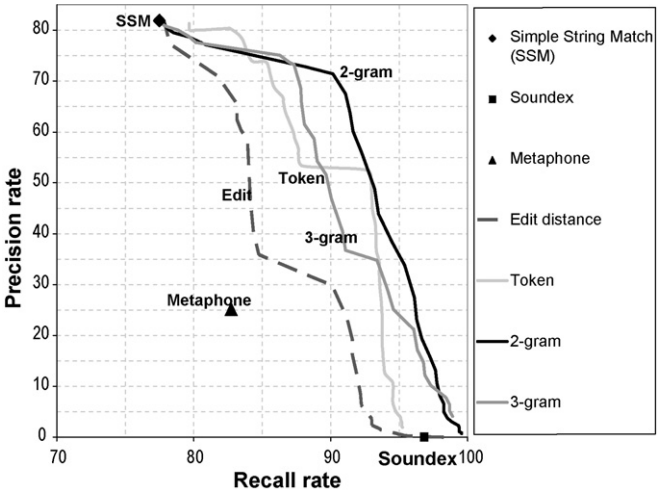


Fig. 2. The impact of different matching algorithms on the precision-recall frontier, using an already multi-parsed string. *Notes*: The comparison among matching algorithms is carried out on multi-parsed data sets. When a *Type I* error occurs, it decreases the *Recall* rate whereas a *Type II* error decreases the *Precision* rate. When applying a matching algorithm we expect an increase in the recalled matches to a given name, which means a decrease in *Type I* errors. But, while the use of these techniques could enlarge the recall rate of misspelled or differently articulated names, it is also expected to enlarge the incorrectly imputed matches, resulting in an increase in *Type II* errors.

and the poorest in terms of recall rate (77%), as it minimizes false positive cases but maximizes false negative ones. From the precision rate point of view, the Token algorithm dominates, providing the same precision rate but with a slight increase in the recall rate (79%). This is explained by the first and last name permutations present in both datasets. However, the decrease in precision starts quickly and the Token algorithm is then dominated by weighted N-gram algorithms when the targeted recall rate is over 82%. Among N-gram algorithms, the 3-gram is found at first sight to be slightly superior when precision is targeted. Below a precision rate of 74% however the 2-gram becomes dominant. Fig. 2 also shows that the usual Soundex algorithm[3], the Edit algorithm and more surprisingly the Metaphone algorithm perform quite poorly and are always dominated by N-gram algorithms.

In order to test the idea that mixed or hybrid algorithms are more efficient than single ones (e.g. Zobel and Dart, 1995; Pfeifer et al., 1996; Hodge and Austin, 2003; Phua et al., 2007), we combined the Token algorithm with both phonetic ones. The Soundex-Token and the Metaphone-Token algorithms – both weighted and unweighted – were tested, resulting in improved performances of the phonetic algorithms. However, these mixed algorithms – not reported in Fig. 2, but available upon request – were still completely dominated by the Token and N-gram algorithms.

When precision is targeted by scholars, the weighted Token algorithm is thus the dominant choice. Instead, when researchers aim at a general identification of a patent portfolio rather than a sampling view and agree to give up some precision in order to reintegrate false negatives, the weighted 2-gram algorithm is a good choice. As shown in Fig. 2, in the latter case, the potential decrease in the precision rate is around 10% whereas the recall rate increases to 13%. In other words, it may be an interesting solution to retrieve 15% of previously neglected patents even if the number of false positive matches rises. This is because scholars can still screen the results afterwards in order to identify the false positives whereas

[2] The particular Soundex function we used here is the one described by Knuth (1973: 391–392), and the particular Metaphone function was the one described by Binstock and Rex (1995). We transformed the Edit distance result dividing it by the maximum length of both text strings and subtracting this result from the unity. The relative similarity reformulation is $1 - [D/\mathrm{Max}(l_1, l_2)]$ where D is the Levenshtein (1966) distance expressed in number of operations and l_i is the number of characters of the text string i. A weighting procedure can be added to Edit transformations, to N-grams and Token vector elements in order to give more less importance to frequent observations (e.g. "street" or "road"). For each gram or Token, we computed the weight as $w_{1i} = 1/(\log n_i + 1)$ and $w_{2i} = 1/n_i$ where n is the number of occurrences in PATSTAT of the token or gram i. N-grams and Token algorithms were then implemented using both w_1 and w_2 weighting vectors and also in a non-weighted fashion. As the latter two were found to be dominated strictly by the use of w_1, they are thus not represented. Results are available upon request.

[3] We used here the original Soundex, which retains only the first four characters. When retaining 8, 16 or 32 characters the recall rate drops steeply while the precision rate does not improve greatly. Results are available upon request.

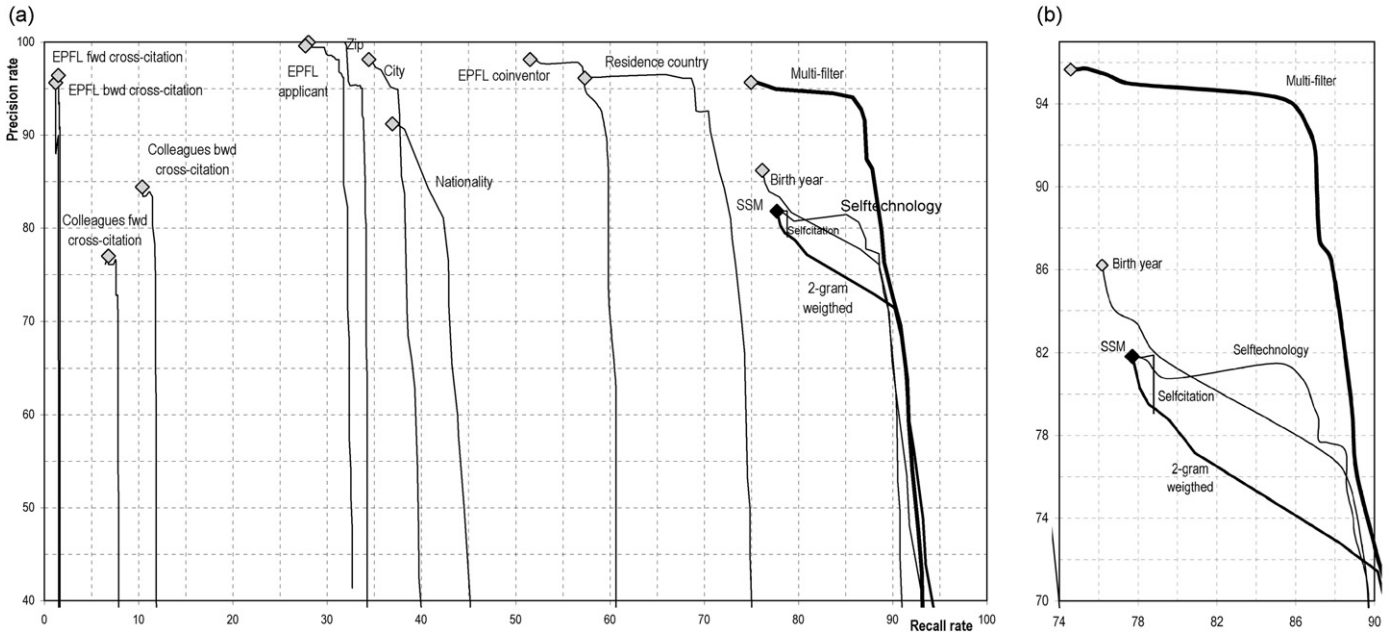


Fig. 3. The impact of different disambiguation filters on the precision-recall frontier, for 2-gram weighted and simple string match algorithms. *Note*: When a single filter is applied, the risk of rejecting correct positives also exists, which would decrease the recall rate. The figure to the right is a zoom of the top right quadrant of the left figure.

false negatives cannot usually be identified after a simple string match. Finally, we also provide evidence that the weighted 2-gram algorithm is more robust than most commonly used algorithms in both precision and recall rates for a wide range of similarity thresholds.

2.3. The filtering stage

The final filtering stage depends on the ability to obtain and implement (through different algorithms or filters) complementary information on individuals in order to identify and reject false positives. Of course, having not only more but also superior information will result in a higher likelihood of improved precision. However, the disambiguation procedures are complex: the use of additional information (Lissoni et al., 2006; Hoisl, 2006) and the ex ante sorting (Hoisl, 2006) or weighting (Trajtenberg et al., 2006) of the available criteria are not straightforward. The relative efficiency of each of these criteria and how to combine them are still open questions.

Complementary information can be divided by its nature into three groups: first, there is the kind of information that is usually simple to obtain when a list of inventors is available (e.g. Mariani and Romanelli, 2007). Besides the names of the inventors, adding their location (e.g. city, region or state) is easy, if it assumed that each inventor lives near the institution he/she is affiliated to. For example, it is easy to compare the residence country, surrounding postal codes and cities where inventors are likely to live against PATSTAT's country and address fields (see "Residence country", "Zip" and "City", respectively, in Fig. 3). Depending on the institution size and multi-localization and accordingly on the number of inventors, it is also easy to find helpful additional information regarding the inventor's field of research or even the different addresses of inventors' units or departments.

A second set of information is also usually easy to obtain through the exploitation of the information available in the patent document itself, such as the citations, co-inventors, applicants or IPC symbols. Given a potential match between a name in the list and the inventor's name in the patent document, the most straightforward use is to check if there is a second (or more) name in the list which matches an inventor's name in the patent document ("EPFL co-inventor" in Fig. 3) or if the inventor's institution appears as an applicant in the same document ("EPFL applicant"). Similarly, as cross-citation is more likely to occur with other inventors from the same institution (see Trajtenberg et al., 2006), the citations of the potentially matched document can be used to verify if they refer to documents where their inventors' names match other names from the list ("Colleagues forward/backward cross-citation") or if the inventor's institution appears as an applicant ("EPFL forward/backward cross-citation"). More complex filters relying on a recursive use of PATSTAT can be also applied. In a nutshell, this consists in validating potential matches by degrees using some measure of certainty and then using the validated matches to compare the remaining potential matches. For instance, after validating a first set of potential matches – ideally with a high precision criterion – the information from those validated pairs of names and patent documents is then used to filter other potential matches. This step can be carried out for example by checking if the potential patent match cites a validated patent of the same inventor ("Self-citations") or if the potential patent has similar IPC symbols to a validated patent from the same inventor ("Self-technology").

The third group is less likely to be obtained as it usually requires access to human resources data – often under confidentiality constraints – or contacting the inventors directly. In our EPFL case, the human resources department provided us with detailed and accurate information on inventors' addresses over time (including their different countries of residence, cities and ZIP codes), nationality, civil status, gender, position and date of birth. Illustrating the third group of filters, the birth year condition is set against the applications filing date using a restriction from 18 to 60 years of active patenting life ("Birth year") and the inventor's nationality ("Nationality") is plotted against the residence country field.

Fig. 3 presents the marginal gains of these different single filters as well as a multiple filter applied after a simple string match and a weighted 2-gram matching process. In general terms, results confirm that filters applied individually may offer some precision gain when compared to the original simple string match and weighted 2-gram, but at a considerable recall loss.

The nature of the complementary information – and the burden of obtaining it – has not a direct influence on the quality of the filter.

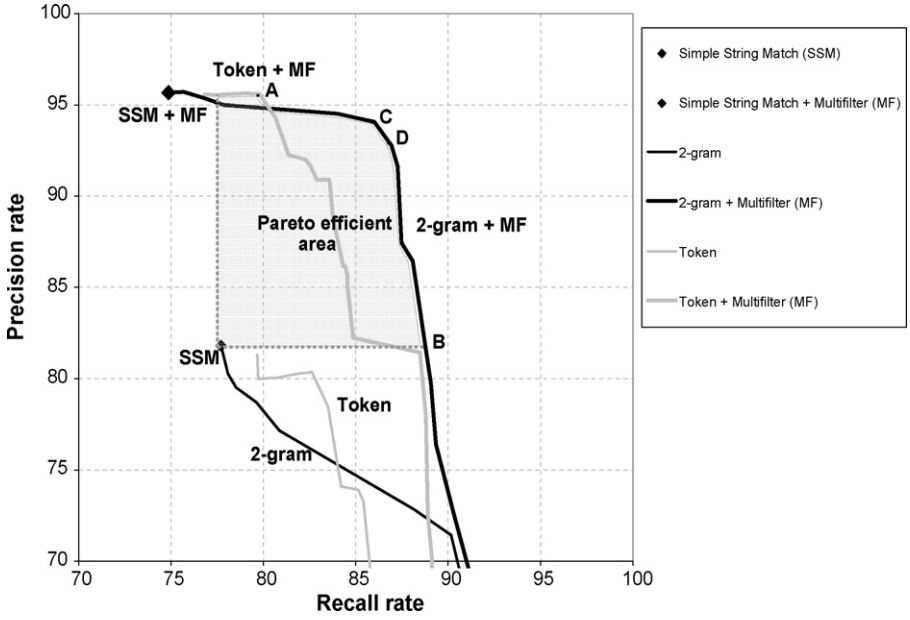


Fig. 4. Recall and precision rates using multiple parsing and disambiguation filters for a weighted 2-gram or for a Token algorithm.

On the one hand, the nationality of inventors, which is much harder to obtain, seems much less effective than the country of residence, this latter being easily inputted. Furthermore, both forward and backward citations are less likely to provide a clear disambiguation than other filters. On the other hand, the inventor's age provides a slight precision improvement with little or even no recall loss. Similarly recursive filters provide some slight improvement when self-citation is considered but much more interesting results are achieved with technology profiles.

However the main result is the gain obtained by applying a multiple filter including all single filters regularly available to scholars (i.e. Birth year, Nationality, Cross-citations and recursive filters are left out). The result of applying such multiple filters with equal weighting is demonstrated at the rightmost side curve of Fig. 3. It is clearly shown that such a disambiguation strategy leads to a significant improvement both in terms of precision (+13%) and recall (+11%)[4]. These results should encourage scholars to be very creative in their quest for additional information, as each sample may respond differently to the filters drawn from the available complementary information. In any case it is the accumulation of filters which will allow scholars to improve their matching techniques in both precision and recall.

3. The complete heuristic and its robustness

3.1. Results

Fig. 4 summarizes the "Names Game" to be played combining the best practices identified in the previous section. The Simple String Match point (SSM) is the matching heuristic usually applied by scholars after the parsing stage and is our benchmark here to assess the gain induced by the use of a more sophisticated matching algorithm combined with a subsequent filtering stage.

Scholars interested in precision will be keen to apply a multi-filter approach after using a simple string match (SSM + MF). The gain is precision is about 14% compared to the SSM only strategy. However, replacing the SSM matching algorithm with the Token algorithm (Point A) is an even better solution since it enables maintaining the highest precision rate while obtaining a recall gain of 5% with respect to a simple string match multi-filtered solution (SSM + MF). On the contrary, scholars aiming to maximize recall can apply a multi-filter after a 2-gram algorithm to minimize the impact on the precision rate. For instance, in point B an improvement of 11% in the recall rate can be achieved without any loss in the precision rate if compared with the simple string match (*ceteris paribus* the parsing stage)[5].

Fig. 4 also suggests that false negatives are likely to be much fewer when one agrees to drop the precision rate slightly. The weighted 2-gram after the multi-filter is quite flat for any similarity threshold between 1 and 0.9. Points C and D in Fig. 4 represent two interesting solutions to consider, respectively the thresholds 0.91 and 0.90.

Therefore, all three stages have to be applied in order to decrease the bias of the matched sample with respect to the real population. At the matching stage, the weighted 2-gram is proof of the best algorithm as it provides a good trade-off between recall and precision. However, it has to be combined with previous multiple parsing and a subsequent disambiguation procedure implementing multiple filters alternatively. By doing this, a large efficiency gain can be achieved in both terms of recovering true matches and discarding false ones. In other words, our results invite scholars to discard SSM in order to apply more effective matching algorithms. Adopting a sophisticated matching algorithm is attractive not only in terms of effectiveness, but also due to the lack of flexibility in the SSM approach, where results are irreversible: although the precision is high, false negatives cannot be retrieved after the matching procedure.

A 2-gram + MF retrieves many more false negatives and, despite the lower precision of a few percent, allows scholars to manually

[4] Adding the remaining filters provided similar results. For instance, adding the birth year filter as a necessary condition improved the multiple filter precision with very little recall loss. Adding technology profile (computed on 4-digit level of the IPC code classification) into the filter generated the recall of false negatives but also of true negatives and thus lowered the precision rate. At the same time it improved the recall rate. Results are available upon request.

[5] However, maximizing the recall rate may have some limitations as the result of applying a Soundex matching algorithm with the same multiple filter gives a high recall rate of 86%, but at the expense of a very low 6% precision rate.

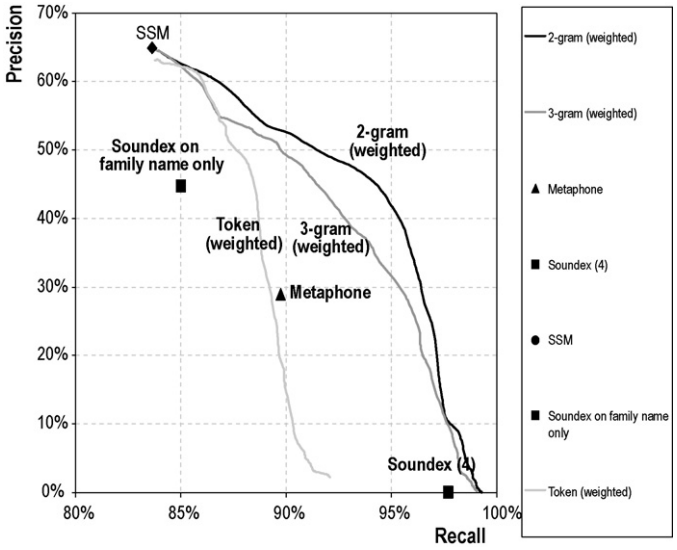


Fig. 5. Recall and precision rates using the Israeli benchmark set.

check after the automatic procedures, in order to get rid of false positives. This last algorithm is thus a more flexible and more interesting option for scholars. An important caveat is however that a similarity threshold (e.g. 0.91) has to be chosen to reduce the list of retrieved pairs to be manually checked. The construction of a small benchmark set to decide which threshold to apply is an initial step to be carried out here before launching the heuristic on the whole list of inventors.

3.2. Robustness

An important question regards the robustness of the 2-gram weighted algorithm as a dominant matching algorithm. Two dimensions are concerned here.

The first deals with the heterogeneity of the list of names. Our results depend on the structure of our benchmark set (Bilenko et al., 2003). Even though the EPFL is a very international school with a large variety of names and surnames, this is not a guarantee that the proposed sequence is reliable for Swedish or Japanese names. In order to test the robustness of our results we decided to apply the different matching algorithms to the list of Israeli inventors used and kindly provided by Trajtenberg et al. (2006). The Israeli benchmark set is a list of 6023 unique Israeli inventors linked to their patents and granted in the US, totaling 15,316 records (see Trajtenberg et al., 2006 for further details).

Fig. 5 reports the results of applying the same multiple parse approach mentioned in the previous section as well as several matching algorithms to the Israeli benchmark sets. In general terms, the results illustrated are in accordance with those presented in Fig. 2, thereby making the case for their robustness. In particular, these results confirm that the weighted 2-gram is the dominant matching algorithm and therefore the one which should be used in the Names Game. The Token algorithm seems to be less appropriate when dealing with the slight spelling differences created when Hebrew names are translated to a Latin form.

A second dimension is the robustness of our results for firms' names. In order to test the robustness of our conclusions, we also applied the same various heuristics to the French list of firms which declared patents in the second French Community Innovation Survey (CIS3). The conclusions were similar to those obtained for inventors' names (Raffo and Lhuillery, 2008, results are available upon request) supporting the hypothesis that the 2-gram weighted matching algorithm is a better solution than the heuristics based on Token, Edit distance, Metaphone, Soundex Simple String Match matching algorithms.

4. Implications for scholars and policy makers

Patent counts are likely to be used by policy makers to approximate invention productivity. The previous section shows that the count produced by simple parsing combined with a simple string match is not the best strategy regarding the identification of errors. One outstanding issue however is to discover if the use of sophisticated methods significantly modifies the results based on the usual simple heuristic (parse + SSM)[6].

4.1. On indicators and descriptive results

The impacts of the different methods can be first assessed on the descriptive works scholars and policy makers often use. We propose here to illustrate the impacts focusing on inventivity, on network description and on technological profiles of the EPFL.

As a first example, we try to account for the patent documents filed by the EPFL inventors in the USPTO and the EPO from 1980 to 2005. Table 1 shows the heterogeneous impact on indicators (e.g. patent counts) of the different heuristics discussed in previous sections, by applying the main strategies deployed in Fig. 4. The differences between the heuristics are important since a gap of 157 patent applications (14% of the total patents in the benchmark set) is observed from the lowest estimation using simple string match with multi-filter (Parse + SSM + MF) to the highest estimation using a weighted 2-gram with the highest recall.[7] Once the burden induced by the use of the maximum recall solution is considered, the different count confirms that the 2-gram algorithm with multi-filter is the better heuristic to minimize the downward bias the names game introduces. In the benchmark set, the EPFL is a patentee in only 252 patent documents, which is only 22% of the patent documents retrieved. Table 1 shows the size of the upward bias which the different heuristics may introduce regarding this indicator. Finally, our results may impact on the analysis of patent value based on citations (Sapsalis et al., 2006). As depicted in Table 1, the citations, both total and average, received by the EPFL portfolio, are usually downward biased and the error is much lower in terms of the heuristics based on 2-grams.

A second implication of our results concerns the mapping of a technological network which is an increasingly popular descriptive tool among scholars and policy makers. The problems due to different heuristics are identified in Fig. 6 which represents the benchmark "ego" network for the EPFL delineated by patent co-applications. Using the same parsing and filtering stages, about 16% of the total networks' nodes (both grey triangles and diamonds in Fig. 6) including either academic bodies (Caltech, Ohio State University, University of Texas) or enterprises (Ericsson, Nec or Genentech) are still unaccounted for by studies relying on precision maximizing algorithms (SSM or Token). It also shows that only 7% of nodes (only grey diamonds in Fig. 6) are unaccounted for by studies relying on a weighted 2-gram heuristic. Furthermore, the heuristics can overlook some closures (in black in Fig. 6 such as the CNRS-Motorola relationships or the University of California-Caltech links but also underestimate the weight of some identified

[6] Note that in many empirical academic papers where a simple string match algorithm is used, it is often argued that the focus on the precision is licit (since the purpose is to identify a portfolio with a minimum of false positives) or that the deletion of false negatives induced by such a matching algorithm should not introduce systematic biases due to a supposed random distribution of these false negative.

[7] These results assume that scholars have manually corrected the false positives after matching, which is a desirable practice. However, dropping this assumption does not change the heterogeneity of outcomes.

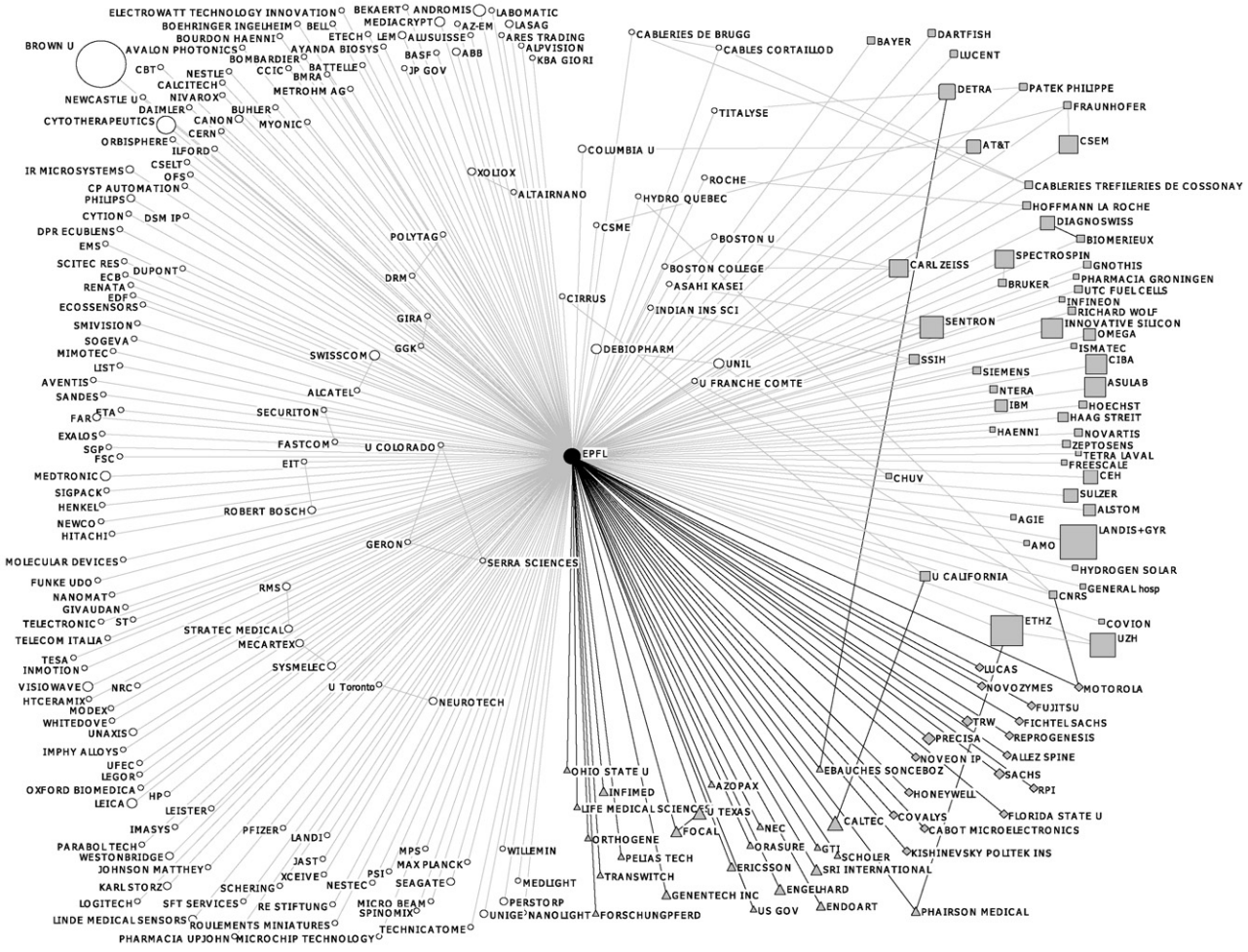
Table 1
The number of retrieved patents and indicators, by algorithms, for EPFL TTO inventors' sets.

	All EPFL TTO's patents[a]		EPFL as applicant	Citations received[b]		Average citations
Parse + SSM	921	80%	25.3%	2470	75%	2.7
Parse + SSM + MF	864	75%	26.9%	2437	74%	2.6
Highest precision (token)	924	81%	25.7%	2558	78%	2.8
Highest recall (2-gram)	1021	89%	23.4%	3106	94%	3.4
Balanced R&P (2-gram)	996	87%	23.7%	3058	93%	3.3
Benchmark set	1147	100%	22.0%	3299	100%	3.6

[a] EPFL Patents are those patent documents filed during the period 1980–2005 with at least one EPFL inventor, where EPFL inventors are those registered by the EPFL's TTO. Patents with unregistered inventors are therefore missing here and consequently the ratios are not representative of the entire inventive activity of the EPFL.
[b] Self-citations were excluded.

nodes (the gray squares in Fig. 6) representing about 20% of total nodes. On average the perfect match strategy neglects 43% of links of these gray squared nodes whereas the 2-gram based heuristic fails to identify only 27% of the links for the same nodes. The implications of this are critical when the complete network is intended to be delineated. Once more, what can be shown is that the 2-gram based heuristic introduces much less downward bias than a single string match based heuristic.

Finally, the third activity often considered by scholars and policy makers is the analysis of the technological profiles of applicants or inventors. The different profiles of the EPFL TTO's portfolio based on the various heuristics' cases are reported in Table 2. Once more,



Notes: Only EPO and US patent documents co-filed during the period 1980-2005 between an EPFL's TTO registered inventor and a non-individual applicant (institutions or firms) were considered. Links represent the co-filing Ego-network of the EPFL's patent portfolio. The size of all nodes represents the real number of co-filed patent documents. Diamonds represent nodes fully unaccounted by both SSM and 2-gram, while triangles represent nodes unaccounted by SSM only and not by 2-gram. Squares represent nodes which are underestimated by SSM or 2-gram. Circles represent nodes fully identified.

Fig. 6. The EPFL co-patenting network, according to different heuristics. *Notes*: Only EPO and US patent documents co-filed during the period 1980–2005 between an EPFL's TTO registered inventor and a non-individual applicant (institutions or firms) were considered. Links represent the co-filing Ego-network of the EPFL's patent portfolio. The size of all nodes represents the real number of co-filed patent documents. Diamonds represent nodes fully unaccounted by both SSM and 2-gram, while triangles represent nodes unaccounted by SSM only and not by 2-gram. Squares represent nodes which are underestimated by SSM or 2-gram. Circles represent nodes fully identified.

Table 2
EPFL's Technological profile, by algorithms.

Benchmark set			Parse + SSM			Parse + SSM + MF			Highest precision[a] (token)			Highest recall[a] (2-gram)			Balanced R&P[a] (2-gram)		
IPC	Rank	%	IPC	Δ rank	%	IPC	Δ rank	%	IPC	Δ rank	%	IPC	Δ rank	%	IPC	Δ rank	%
A61K	1	8.0%	A61K	0	7.0%	A61K	0	7.4%	A61K	0	7.0%	A61K	0	8.2%	A61K	0	8.3%
H01L	2	5.2%	H01L	0	5.5%	H01L	0	5.4%	H01L	0	5.7%	H01L	0	5.3%	H01L	0	5.1%
A61L	3	4.9%	G01N	−1	4.5%	C12N	−2	4.5%	C12N	−2	4.3%	G01N	−1	4.6%	A61L	0	4.7%
G01N	4	4.6%	C12N	−1	4.3%	G01N	0	4.3%	G01N	0	4.1%	A61L	1	4.6%	G01N	0	4.7%
C12N	5	4.0%	G01R	−1	3.8%	G01R	−1	3.7%	G01R	−1	3.8%	C12N	0	4.1%	C12N	0	4.2%
G01R	6	3.3%	A61F	−2	3.4%	H04N	−3	3.5%	H04N	−3	3.5%	G01R	0	3.4%	H04N	−3	3.4%
A61B	7	3.1%	A61L	4	3.3%	A61L	4	3.3%	A61L	4	3.2%	H04N	−2	3.4%	G01R	1	3.4%
A61F	8	3.1%	H04N	−1	3.3%	A61F	0	3.3%	A61F	0	3.2%	A61F	0	3.1%	A61F	0	3.2%
H04N	9	3.0%	A61B	2	3.0%	B01J	−1	2.8%	A61B	2	2.9%	B01J	−1	2.8%	B01J	−1	2.9%

[a] With parsing and disambiguation stages applied.

important differences are highlighted regarding the relative importance of different technologies within the same patent portfolio. For example, a standard heuristic (parse + SSM) underestimates the importance of IPC technological fields such as "Methods or apparatus for sterilizing materials or objects in general (A61L)" which is pooled with the main and complementary technologies related to "medical preparations (A61K)."

The different statistics produced here underline the fact that the analysis and comparisons drawn on matched data on PROs, firms or even countries, can be seriously misleading for scholars and policy makers. The choice of an inefficient algorithm introduces significant biases in the collection of patent data leading to critical differences in patent count and value, network representations and technological profiles.

4.2. On micro-econometric results

Beyond descriptive statistics, the biases introduced by the heuristics may affect econometric results. In order to test if the studies exploring the propensity of academic employees to patent

Table 3
Negative Binomial regression on inventors' patents (USPTO and EPO), according to the matching strategy.

	(1) Benchmark ("real")	(2) Only parsed (P + SSM)	(3) Parsed & filtered (P + SSM + F)	(4) Highest precision (P + Token + F)	(5) Balanced R&P (P + 2-gram + F)	(6) Highest recall (P + 2-gram + F)
Age	0.292***	0.234***	0.200***	0.232***	0.225***	0.233***
	(0.066)	(0.068)	(0.062)	(0.061)	(0.061)	(0.058)
Age squared	−0.003***	−0.003***	−0.002***	−0.003***	−0.003***	−0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	−0.194	−0.308	0.196	0.277	0.112	−0.099
	(0.534)	(0.552)	(0.393)	(0.380)	(0.391)	(0.385)
Single	−0.091	−0.004	−0.232	−0.233	−0.271	−0.133
	(0.254)	(0.258)	(0.217)	(0.214)	(0.212)	(0.200)
Non-Swiss national	−0.558**	−0.654**	−0.550***	−0.556***	−0.564***	−0.493***
	(0.253)	(0.259)	(0.205)	(0.201)	(0.197)	(0.188)
Professor (tenured)	0.387	0.412	0.553*	0.487*	0.411	0.330
	(0.327)	(0.347)	(0.297)	(0.291)	(0.286)	(0.282)
Professor (not tenured)	0.044	0.036	0.214	0.118	0.264	0.127
	(0.228)	(0.245)	(0.212)	(0.208)	(0.199)	(0.194)
Foreign Professor (tenured)	0.704*	0.340	0.181	0.437	0.576*	0.547
	(0.385)	(0.400)	(0.359)	(0.352)	(0.345)	(0.338)
Field: Architecture	0.355	0.408	0.881**	1.121***	0.760**	0.794**
	(0.359)	(0.417)	(0.364)	(0.348)	(0.332)	(0.324)
Field: Basic Sciences	0.627**	0.803***	0.360*	0.273	0.268	0.160
	(0.259)	(0.270)	(0.202)	(0.197)	(0.195)	(0.188)
Field: Computer Sciences	0.128	0.205	0.061	0.027	0.007	−0.187
	(0.346)	(0.358)	(0.275)	(0.273)	(0.271)	(0.267)
Field: Others	−0.368	−0.454	−0.459	−0.449	−0.207	−0.325
	(0.321)	(0.384)	(0.358)	(0.342)	(0.307)	(0.304)
Constant	−6.650***	−5.632***	−6.235***	−6.833***	−6.583***	−6.515***
	(1.445)	(1.482)	(1.353)	(1.332)	(1.315)	(1.254)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Inventor-years	2050	1887	1856	1886	1962	1985
Number of Inventors	267	246	242	246	254	258
Log-Likelihood	−1195.00	−1080.20	−1339.95	−1397.48	−1457.04	−1558.16
Chi squared	51.01***	41.22***	50.27***	65.37***	59.11***	54.48***

Notes: Standard errors reported in parenthesis; P = Multiple Parse, F = Multiple Filter, SSM = Simple String Match, R&P = Recall and Precision.
Foreign Professor (not tenured) is not included since the variable is highly collinear with that of Professor (not tenured) variable.
Similarly, experience at the EPFL is also removed since it is very collinear to age.
Life science inventors are removed from the sample since the faculty only opened recently (2003).
Engineering science is taken as a reference.
Note that the results in column (5) and (6) were obtained without any manual check. Such a step should make the results converge toward the results displayed in column (1).

*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

are reliable, and thus to know if it is worth performing the more sophisticated matching procedure we are proposing here, we built a count data model on the different patent counts found for EPFL inventors. The count data model with fixed effects[8] is estimated 6 times, and implements the following various variables: Model (1) explains the actual patent count from our benchmark set. This model is considered as the real model to be compared with. Model (2) regresses the patents matched using a simple string match after multiple parsing. Model (3) uses the patents matched using a simple string match combined with a multiple parsing and a multiple filter. Model (4) considers the patents matched aiming at the highest precision rate (point A at Fig. 4). Model (5) implements the patents matched aiming at a balance between recall and precision (Point C at Fig. 4). Finally, model (6) employs the patents matched aiming for the highest recall without losing precision with respect to the simple string match (Point B at Fig. 4).

Results reported in Table 3 (see the Appendix for econometric details) suggest that the use of the various algorithms leads to different results. While the influence of several variables is found to be robust (age, foreign nationality), others, such as career level variables or scientific field are found to be heterogeneous in both coefficient and significance. The heuristic which deals with both Precision and Recall is the only one which correctly identifies the positive effect of tenured foreign professors. Heuristics focusing only on precision suggest that tenured professors are more likely to patent which is not true according to our benchmark set. Note that the results obtained on maximized recall are found to be reasonably correct.

The volatility of results underlines the importance of the "Names Game" stage in patent based research. The results cast doubt on the relevance of precision as a target in the names game and thus suggest a lack of reliability of econometric results obtained with algorithms based on simple string match or token algorithms. The results also show that solutions exist in order to mitigate the problem. The adoption of a 2-gram algorithm combined with manual checking of proposed pairs seems to be a workable solution.

5. Conclusion

Patent data are now extensively used by scholars. Furthermore, patent data are increasingly matched with other lists of firms' or inventors' names in order to rebuild patent portfolios. Little is known however about the reliability of the procedures implemented by researchers. A breakthrough paper by Trajtenberg et al. (2006) unveiled the critical importance and the complexity of this names game and proposed a single heuristic. However, we still do not know what the best heuristics to apply are when using different parsing, matching and filtering stages.

Using the PATSTAT dataset and an EPFL benchmark list of inventors, we explored several solutions. The first general result is that parsing, matching and filtering are three important stages whose interactions are so important that in order to achieve high precision and recall rates, no relaxation in the implementation of any one of them can be afforded. The present paper also shows that some heuristics are more reliable than others. The use of the 2-gram algorithm in the matching stage with complete parsing and filters provided the opportunity to directly retrieve a high number of patents, limiting the number of false positives to be searched manually, but also retrieving false negatives usually neglected. Our solution enabled us to get rid of up to 78% of false positives, and to rehabilitate up to 45% of false negatives, both of which are typically unsolved by the perfect string match implemented by scholars. All in all, 50% of the errors usually unaccounted by the perfect string match are corrected for when using the complete heuristic we propose as a solution. A further interesting result is that the dominance of the 2-gram heuristic result seems robust over a large range of names and surnames including firms' names.

The use of such a sophisticated heuristic facilitates a more accurate, if not perfect, view of patent portfolios of agents or institutions. It can therefore significantly modify the results on inventive activity, technological profiles or networking and modify the micro-econometric results obtained, thereby explaining patent count for example. The present article shows that without the use of an efficient heuristic, the reliability of econometric results and of S&T descriptive statistics is questionable. A lack of a convenient matching heuristic may induce false academic results, erroneous strategies for firms or inaccurate decisions for policy makers. With respect to the comparisons between two results based on the names games applied to different lists of inventors (or firms), our results suggest that the observed differences or similarities among inventors, institutions, firms or countries can even be due to the adoption of different matching heuristics, including different initial information on inventors (or firms). It also shows that even when the matching methodology implemented is clearly identified and homogenous among the studies, it does not offer any guarantee of accuracy as the applied heuristic is awkwardly chosen.

One problem is that the proposed solutions are cumbersome since they demand a lot of computational load and require the use of all national patent data available in PATSTAT. The present paper presents the first workable, albeit not perfect, solution. Further research should improve this heuristic in many ways: in the matching stage, few mixed algorithms were explored in the present paper without achieving the better results they were supposed to. Solutions superior to that of using a lone 2-gram algorithm could however be found by further investigation. The same remark applies for the other stages. For example, the identification and verification of proper addresses for applicants or inventors is an interesting strategy to improve the parsing stage. Similarly interesting gains could be obtained comparing the information on patents belonging to the same patent family. The comparison of abstracts contained in articles and publications (à la Cassiman et al., 2007) could also be a complementary tool in order to achieve a better disambiguation stage. A further aspect of this is that the proposed heuristic, robust as it is, is limited in is scope since it does not properly address all languages. For example, the identification of Japanese or Chinese patents may be problematic as either matching with the applicant's name can be difficult[9] or the disambiguation stage becomes essential (e.g. very frequently, inventors have short surnames such as Wang).

We contend however that the main caveat of the proposed solution is that it is likely to be useless if not applied systematically. The utility associated with the proposed solution is indeed associated with its general adoption by scholars, statistical offices and policy makers. The more the same reliable rules are adopted, the more the results will be consistent and comparable. Without such a systematic adoption, statisticians and researchers will either be trapped into the production of false results due to the inefficiency of their

[8] A negative binomial model for panel data with fixed effects is an extension of the Poisson regression model which allows the variance of the process to differ from the mean. The explanatory variables used in the empirical model are listed in Table 2 and are standard in the literature on academic patenting (see, for example, Azoulay et al., 2007).

[9] Despite the fact that applicants are now asked to fill in the form in their original language, applicant names are often found to be translated (before 1989, only Katakana characters were accepted). The problems arise for non-Japanese firms (e.g. "FUORUKUSUWAAGEN" is the phonetic translation found for VOLKSWAGEN and "JIIMENSU" the one for SIEMENS) or even for Japanese ones (e.g. "SONII" can be found as an applicant as a phonetic translation of SONY).

basic heuristic (e.g. single string match) or confronted with flawed comparisons based on heterogeneous "home made" heuristics. We hope that the availability of the different algorithms used in this paper will accelerate the expected convergence[10].

Acknowledgements

We gratefully acknowledge the EPFL TTO and the EPFL Human Resources for data availability. When producing this paper, we also benefited considerably from the expertise of several computer scientists and engineers. We especially want to acknowledge the collaboration of A. Gonçalves (UFSC), L. Matas (UBA/CAICYT) and F. Lladós (ITBA/UPM). The latter was of great help when implementing the N-grams algorithms and he is primarily responsible for the development of its high-performing indexation. Many thanks also to M. Trajtenberg, G. Shiff and R. Melamed for kindly providing us with their data on Israeli inventors. Finally, we also acknowledge D. Guellec, C. Le Bas, C. Martinez, J. Rollinson, A. Schoen, G. Thomas, N. Van Zeebroeck for their comments and the participants at the EPO Conference on Patent Statistics for Policy Decision Making, Venice International University, San Servolo, Venice, Italy, 1–3 October 2007, and the attendants at the 3rd Annual Conference of the EPIP Association, Bern, Switzerland, October 3rd–4th, 2008. Any errors are our own.

Appendix A. The different matching algorithms

Matching algorithms likely to be useful in the second stage of the names game can be chosen among three main families: Phonetic algorithms, Edit distance algorithms and vectorial decomposition algorithms. First, the Phonetic algorithms (such as Soundex, Daitch-Mokotoff Soundex, NYSIIS, Double Metaphone, Caverphone, Caverphone 2.0, Phonix, Onca, Fuzzy Soundex, etc.) regroup phonemes by sound proximity using a simple set of rules. The Soundex algorithm for example transforms the string '*Aebischer Patrick*' or '*Abisher Patric*' into a single A126. Metaphone will transform both into the same EBSXRPTRK. A simple string match between these transformed strings is thus performed. Second, the Edit distance algorithms (such as Levenshtein distance, Damerau–Levenshtein distance, Bitap, Hamming distance, Boyer-Moore, etc.) are also based on a simple precept, which is that any text string can be transformed into another by applying a given number of plain operations. Transforming '*Abischer Patric*' into '*Aebischer Patrick*' requires for example only 2 insertions. The algorithm provides lists of potential matching where positive matches are likely to be those with the lowest number of transformations. Finally, the family of *vectorial decomposition algorithms* (such as 2-Gram, 3-Gram, 4-Gram, Token, etc.) is basically a comparison of the elements of both strings. The N-gram algorithm decomposes the text string into elements of N characters, called grams, on a moving windows basis. For example, a 3-gram decomposition of *Aebischer Patrick* will include 15 3-grams: AEB, EBI, BIS, ISC, SCH, CHE, HER, ER_, R_P, _PA, PAT, ATR, TRI, RIC and ICK. When compared with the name *Abischer Patric* for example, this pair shares only 9 trigrams. The Token algorithm splits the text string by its blank spaces into different elements, called tokens. In our example, '*Aebischer*' and '*Patrick*' are the only two tokens identified. Once both compared text strings are decomposed into elements, a similarity indicator can be computed by applying the cosine distance between both vectors of elements (either grams or tokens) or any other measure. Positive matches are likely to be inventors with the highest cosine distance.

Each algorithm has its own merits. For example, phonetic based algorithms are more efficient at managing similar sounds based on misspellings. The Edit distance algorithm family manages typing or spelling errors effectively. The N-gram algorithms work effectively on misspellings as well as large string permutations. The different algorithms are however usually customized to improve their performances. For example, a weighting procedure can be added to the Edit transformations or to the N-grams and Token vector elements in order to give more importance to observations or changes that are less likely to occur in a text. In N-gram or Token algorithms, grams or tokens (e.g. "street" or "road") which are more present in inventors names and addresses provide less information than rare grams or tokens. A simple approach therefore is to weight grams or tokens assigning them a weight equal to the inverse number of their occurrences in the database.

Several rankings of matching algorithms are already available in the literature on name matching (see Pfeifer et al., 1996; Zobel and Dart, 1995; Phua et al., 2007). Even though a clear hierarchy is hard to achieve for several reasons, Phonex or 2-gram are found to be better performers than 3-gram, 4-gram, or Damerau–Levenshtein algorithms (Pfeifer et al., 1996; Phua et al., 2007; Christen, 2006). According to the surveyed literature, hybrid matching algorithms have even better results (e.g. Zobel and Dart, 1995; Pfeifer et al., 1996; Hodge and Austin, 2003; Phua et al., 2007). Although we have explored the outcomes of some of the various kinds of complex matching algorithms, their systematic examination is beyond the scope of this paper.

Appendix B. The data sources and benchmark set

The datasets we used to perform the tests on matching algorithms are from three different sources. The first and largest is the September 2006 version of "PATSTAT" (the European Patent Office (EPO) Worldwide Patent Statistical Database). It contains approximately 12 million inventors' names who filed a patent application at either the EPO or the United States Patent and Trademark Office (USPTO).

The EPFL Technology Transfer Office (TTO) provided a list of 841 inventors listed in 1995–2005 EPFL patents. This list contains the 1995–2005 EPFL inventors, defined as any inventor(s) or co-inventor(s) of a declared invention made at the EPFL. These EPFL inventors were more likely to be registered in EPO or USPTO records over the same period[11] as most EPFL declared inventions are patented. The TTO list provides additional information on inventors: name, surname, middle name, ZIP codes of personal addresses, their scientific research laboratory, co-inventors with the ZIP code of the personal address, and lastly co-ownership.

Finally, the TTO information on these inventors was completed with the data from the 1994–2006 EPFL employee register provided by the EPFL human resources department. This register provided us with 8885 non-administrative different EPFL names. Among the 841 EPFL inventors, 515 were found to be employed for at least 1 year over the period. The annual Human Resource list for employees includes names, surnames, middle names (not systematic), gender and also the different ZIP codes of personal addresses over the period. Thanks to one's birth date or the EPFL individual code, the human resources list also provides a clear distinction between homonymic researchers and potential changes in names. Changes in family names due to marriages or divorces should be scarce as the EPFL is an engineering school where female teachers or researchers are still a minority.

[10] The scripts in SQL or Php can be downloaded on http://cemi.epfl.ch/ where additional information on PATSTAT users is provided.

[11] We acknowledge that these inventors may also file their inventions at other patent offices, but this is unlikely to happen without equivalent filing at the EPO or USPTO.

Not all patent documents before 2005 are available however and some inventions are not protected by patents (software for example). The information on inventors and patent documents was complied both manually and by applying the different filters and algorithms listed previously. The resulting large list of matches was checked manually one by one for possible errors. We succeeded in identifying 374 EPFL employees out of 515, having at least one patent document in PATSTAT (USPTO and EPO only), meaning a total of 2607 pairs of names. After this first step, around 777 pairs remained ambiguous. This is because an EPFL inventor can put the address of the institution where she or he is a visiting professor for 6 months, but there is no way of verifying whether the two inventors' names refer to the one person or not, except by direct contact. We decided to clean our benchmark set for these ambiguities. Not continuing with these 777 pairs allowed us to keep a clear distinction in our results between what is available without complex heuristics (i.e. using a perfect match) and what cannot be solved by such heuristics without additional information, something which is usually not available from existing datasets.

In conclusion, our core benchmark set was composed of 1830 pairs representing 349 EPFL researchers and their patents filed in EPO or USPTO. The benchmark data set was then used in order to assess the performances of different heuristics.

References

Azoulay, P., Ding, W., Stuart, T., 2007. The determinants of faculty patenting behavior: demographics or opportunities? Journal of Economic Behavior & Organization 63 (4), 599–623.

Balconi, M., Breschi, S., Lissoni, F., 2004. Networks of inventors and the role of academia: an exploration of Italian patent data. Research Policy 33 (1), 127–145.

Bilenko, M., Mooney, R.J., Cohen, W.W., Ravikumar, P., Fienberg, S.E., 2003. Adaptive name matching in information integration. IEEE Intelligent Systems 18 (5), 16–23.

Binstock, A., Rex, J., 1995. Practical Algorithms for Programmers. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

Cantner, U., Graf, H., 2006. The network of innovators in Jena: an application of social network analysis. Research Policy 35 (4), 463–480.

Cassiman, B., Glenisson, P., Van Looy, B., 2007. Measuring industry-science links through inventor-author relations: a profiling methodology. Scientometrics 70 (2), 379–391.

Christen, P., 2006. A comparison of personal name matching: techniques and practical issues. In: Proceedings of the Workshop on Mining Complex Data (MCD). IEEE International Conference on Data Mining (ICDM), Hong Kong, December.

Freeman, C., Soete, L., 2009. Developing science, technology and innovation indicators: what we can learn from the past. Research Policy 38 (4), 583–589.

Hodge, V.J., Austin, J., 2003. A comparison of standard spell checking algorithms and a novel binary neural approach. IEEE Transactions on Knowledge and Data Engineering 15 (5), 1073–1081.

Hoisl, K., 2006. German PatVal Inventors – Report on Name and Address-Matching Procedure. Unpublished manuscript. University of Munich. http://www.inno-tec.bwl.uni-muenchen.de/files/forschung/publikationen/hoisl/patval_matching.pdf.

Kim, J., Lee, S.L., Marschke, G., 2005. The influence of university research on industrial innovation. NBER Working Papers 11447. National Bureau of Economic Research, Inc.

Knuth, D., 1973. The Art of Computer Programming, vol. 3: Sorting and Searching. Addison-Wesley, pp. 391–392.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8), 707–710.

Lissoni, F., Sanditov, B., Tarasconi, G., 2006. The KEINS Database on Academic Inventors: Methodology and Contents. WP N.181, September.

Magerman, T., Van Looy, B., Song, X., 2006. Data production methods for harmonised patent statistics: patentee name harmonization. K.U. Leuven FETEW MSI Research Report 0605, Leuven, March, 88p.

Mariani, M., Romanelli, M., 2007. "Stacking" and "picking" inventions: the patenting behavior of European inventors. Research Policy 36 (8), 1128–1142.

OECD, 2002. Frascati Manual, Proposed Standard Practice for Surveys on Research and Experimental Development. OECD Publishing, Paris.

Pfeifer, U., Poersch, T., Fuhr, N., 1996. Retrieval effectiveness of proper name search methods. Information Processing & Management 32 (6), 667–679.

Phua, C., Lee, V., Smith-Miles, K., 2007. The personal name problem and a recommended data mining solution. In: Encyclopedia of Data Warehousing and Mining, 2nd ed. IDEA Group Publishing.

Raffo, J., Lhuillery, S., 2008. Matching procedures and harmonization methods, Workshop on Harmonisation of Applicants' names in Patent Data. OECD, 13th of March, Paris La Defense.

Sapsalis, E., van Pottelsberghe de la Potterie, B., Navon, R., 2006. Academic vs. industry patenting—an in-depth analysis of what determines patent value. Research Policy 35(10) 1631–1645.

Thoma, G., Torrisi, S., 2007. Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases, CESPRI Working Papers 211. CESPRI, Centre for Research on Innovation and Internationalisation, Universita' Bocconi, Milano, Italy, revised December.

Thursby, J., Fuller, A.W., Thursby, M., 2009. US faculty patenting: inside and outside the university. Research Policy 38 (1), 14–25.

Trajtenberg, M., Shiff, G., Melamed, R., 2006. The "Names Game": Harnessing Inventors' Patent Data for Economic Research. NBER working paper series, No. w12479, September.

Zobel, J., Dart, P., 1995. Finding approximate matches in large lexicons. Software – Practice and Experience 25 (3), 331–345.