Who's Who in Patents. A Bayesian approach

Nicolas Carayol [‡,1], Lorenzo Cassi[◊]
[‡] GREThA, Université Bordeaux IV - CNRS
[◊] CES, Université Paris 1 Panthéon Sorbonne - CNRS

This version: June 2009

[1]Corresponding author. Nicolas Carayol, Université Bordeaux IV, GREThA - CNRS, Avenue Leon Duguit, F-33608 Pessac Cedex. Tel: +33-556848640. Email: nicolas.carayol@u-bordeaux4.fr

Abstract

This paper proposes a bayesian methodology to treat the who's who problem arising in individiual level data sets such as patent data. We assess the usefullness of this methodology on the set of all French inventors appearing on EPO applications from 1978 to 2003.

Keywords: Patents, homonymy, Bayes rule

JEL codes: C81, C88, O30

1 Introduction

The seminal contributions of Schmookler (1966), Scherer (1982) and Griliches (1984) clearly pointed out the relevance of patent data for economic analysis. Since then, the successive waves of empirical contributions in the domain seem to have mainly relied upon access to new information produced from patent data. For instance, the series of articles in Jaffe and Trajtenberg (2002) build upon the access to the USPTO applications data on citations to patent literature. A new challenge for economic analysis obviously resides in the reliable identification of inventors in patent data. Such information could allow the profession to deeply revisit the investigation of knowledge flows, either through network analysis (see e.g. Jackson and Rogers, 2007 for a general analysis of network structures and Carayol and Roux, 2008 for a first investigation of the structure of co-invention networks) or through the systematic investigation of inventors' mobility (in space or across assignees). Nevertheless that issue is not trivial since we face a large scale "who's who" issue due to the homonymy of inventors and spelling errors. Most of the time such errors cannot be neglected since small identity errors usually make great changes in the data. For instance, a few (positive) homonymy errors lead to consider that two different persons are the same. Thus, one mistakenly generate some very connected agents who will abusively link different communities (the ones to which the "true" agents are connected). Negative homonymy errors would lead to the opposite, that is to ignore the role of bridging agents. As it is well known in the literature on networks, statistics describing the network such as the average inter-individual distance (to which the effectiveness of knowledge diffusion may be associated) are significantly affected by such errors. Therefore, the use of the information on patent inventors requires the correct identification of individual identities in patent data through some reliable, systematic and reproducible methodology.

A series of large scale studies already intended to tackle the issue of the identification of inventors identities has adopted more or less some ad hoc techniques. Singh (2005) assumes an inventor can be fully identified by an identical first, last name, and middle initial (when the latter was blank, the technological subcategories need tooverlap). Fleming et al. (2007) rely on the frequencies of last names and the overlap of co-inventors. The most systematic contributions to the issue are the ones of Lissoni et al. (2006) and Melamed et al. (2006) which provide comprehensive techniques for matching inventors with same names and

first names in the former article and with different names using the Soundex system in the second one. In both papers, pairs of potentially identical inventors are matched on the basis of a similarity scoring relying on information on location, assignee, technological classification, citations, and the overlap of co-inventors. Nevertheless their techniques are not based on clear theoretical grounds and thus face two major drawbacks. First, they arbitrarily assign a given increase in the similarity scores between two identities when recording that they have the same observation for some variable. For instance, why should living in the same city count more or less than patenting in the same technological class to infer that two identities correspond to the same person? Second, the relative frequencies of each variable modality are not taken into account. For instance, positive homonymy errors are more likely for persons having frequent names. Similarly, two homonyms are more likely to be different persons if they are both localized in a large city rather than if they are both localized in a small town.

This paper introduces a Bayesian methodology for estimating that two *ex ante* different persons are the same given a series of observables provided by the data.[1] This methodology fully overcomes the two drawbacks of previous studies stressed above. It is applied on a dataset of all inventors listed in European patent applications having an address in France from 1978 to 2003, which implies nearly $237,000$ inventor$\times$patent occurrences. Given that an empirical benchmark is available (a relatively limited list of undoubtable observations on potentially identical agents), we can determine a threshold value which minimizes any linear combination of positive and negative errors in the benchmark data. Our main empirical result is that the minimal weighted share of negative and positive errors in our preferred specification is less than two percents over potential ones.

The Bayesian methodology in developed in Section 2 while the estimations and the results are presented Section 3. Section 4 concludes.

2 Theory

Let us first consider a list I of identities or *ex ante* agents i, that is agents for whom we surely (or almost surely) know that, though we may observe them several times in

[1] Though this methodology is developed for inventors in patent data, it can be applied to other data having a similar structure.

the data, they are the same person. Each *ex ante* agent i is characterized by a series of K variables (including name and first name) labeled X^k (in vector form $1 \times \#I$), with $k = 1, ..., K$. We assume that these variables are mutually independently distributed, an hypothesis that simplifies considerably the exposure while it may be relaxed easily.

The main goal of this paper is to provide a methodology to estimate the probability that any *ex ante* agent $i \in I$ is the same person than some other *ex ante* agent $j \in I$. In short, we are in search of the partition $\pi = \{C_1, ..., C_m\}$ of I, with $\cup_{i=1,...,m} C_i = I$, and $C_j \cap C_h = \emptyset, \forall j, h = 1, ..., m$ with $j \neq h$, which corresponds to the "correct" *ex post* identities. We note $\{i, j\} \subset C_h$ by writing "$i = j$".

In order to assess the probability of that event, we may condition on the observables on i and j, that are draws of random variables X_i^k and X_j^k for all $k = 1, ..., K$ and their respective frequencies of occurrence. Without loss of generality, let's assume that for some *ex ante* agents i and j, we have $X_i^k = X_j^k$, for all $k = 1, ..., \bar{k} - 1$ and $X_i^k \neq X_j^k$ for all $k = \bar{k}..., K$. Thus, we intend to estimate the following conditional probability:

$$\Pr\left(i = j \left| X_i^k = X_j^k, \forall k = 1, ...\bar{k} - 1 \text{ and } X_i^{k'} \neq X_j^{k'}, \forall k' = \bar{k}, ...K \right.\right). \tag{1}$$

That probability shall in turn ground the calculation of *similarity scores* between i and j on which decisions relative to who's who shall be taken.

2.1 A Bayesian approach

To calculate a variation of (1), we rely upon a Bayesian approach. Before, we need first to compute the probabilities that *ex post* (real) agents change their observables from one invention occurrence to the other. Let's write ε^k the probability that any *ex post* agent changes his k^{th} variable between two invention occurrences in which he was identified for instance as *ex ante* agent i and *ex ante* agent j: $\varepsilon^k \equiv \Pr\left(X_i^k \neq X_j^k, |i = j\right)$. We further

assume that ε^k is independent of $\varepsilon^{k'}, \forall k \neq k'$.[2,3] Therefore, we can write:

$$\begin{aligned} &\Pr\left(X_i^k = X_j^k, \forall k = 1, ...\bar{k}-1, \text{ and } X_i^{k'} \neq X_j^{k'}, \forall k' = \bar{k}, ...K \middle| i = j\right) \\ &= \Pi_{k=1,...\bar{k}-1}\left(1-\varepsilon^k\right)\Pi_{k'=\bar{k},...,K}\varepsilon^{k'}. \end{aligned} \tag{2}$$

We can now apply the Bayes rule, according to which the probability in (1) is equal to

$$\frac{\Pr(i=j) \times \Pr\left(X_i^k = X_j^k, \forall k = 1, ...\bar{k}-1, \text{ and } X_i^{k'} \neq X_j^{k'}, \forall k' = \bar{k}, ...K \,|i = j\right)}{\Pr\left(X_i^k = X_j^k, \forall k = 1, ...\bar{k}, \text{ and } X_i^{k'} \neq X_j^{k'}, \forall k' = \bar{k}, ...K\right)}.$$

Since it is not possible to compute $\Pr(i=j)$, we focus on $\Delta(i,j)$ the increase in the probability that i and j knowing that $X_i^k = X_j^k, \forall k = 1, ...\bar{k}-1$ as compared to knowing the reverse. Using (2), and after some computations and combinations, it comes:

$$\Delta(i,j) = \Pi_{k=1,...\bar{k}-1} \quad \beta^k \times \Omega^k(i,j) \tag{3}$$

with $\beta^k \equiv \frac{\left(1-\varepsilon^k\right)}{\varepsilon^k}$ and $\Omega^k(i,j) \equiv \frac{\left(1-\Pr\left(X_i^k = X_j^k\right)\right)}{\Pr\left(X_i^k = X_j^k\right)}$. β^k is the probability that any agent has two different values of her k^{th} variable through any two different identities she may take divided by the reverse. $\Omega^k(i,j)$ is the probability that the two *ex ante* agents i and j have different observables divided by the reverse (irrespective to the fact that they are or are not the same persons *ex post*). This term accounts for the frequency of occurrence of the observables (having the same observable counts more when this observable is less frequent). The similarity score thus fully integrates endogenously the frequency of occurrence of the observables on i an j when they are identical through $\Omega^k(i,j)$ and the propensity of *ex ante* agents to change each of these variables through β^k.

2.2 Thresholds and the transitivity of identities

Once the similarity index is defined, a threshold value of the similarity has to be established. Below such threshold, two *ex ante* agents are to be declared as different agents

[2]This assumption, which has been introduced to simplify the exposure, could be relaxed easily. It means that for instance the probability that someone changes his address from one invention to the other is independent of the probability this person changes from one technological field to the other.

[3]For some variables (name, first name without spelling errors) we are inclined to think that $\varepsilon^k = 0$, whereas for some others (address, technological field...), clearly $\varepsilon^k > 0$.

and above it they are assumed to be the same person *ex post*. As soon as such a threshold $\bar{\Delta}$ is defined, a transitivity issue arises.[4] For instance, consider three *ex ante* agents z, w and h and $\Delta(z,w) > \Delta(z,h) > \bar{\Delta} > \Delta(h,w)$. Then, *ex ante* agents z and w are *ex post* considered as being the same person as well as z and h. If these two statements apply then obviously, h and w are also the same person *ex post* by transitivity. That is if $z, w \in C_h$ and $z, h \in C_j$, then necessarily $h = j$ since $C_j \cap C_h = \emptyset, \forall j \neq h$.

We thus need to improve the values of similarity scores $\Delta(i,j)$ so as to take into account that transitivity of identities. To do so, the following algorithm is proposed. It recursively upgrades similarity scores when a transitive triplet of identities suggests so and does so until one can not find anymore such a configuration in the data.

Algorithm 1 *For all considered pairs of distinct ex ante agents i and j, we apply:*

$$\Delta(i,j) \hookleftarrow \max\left(\Delta(i,j); \max_{h \in I \setminus \{i,j\}} \min(\Delta(i,h); \Delta(j,h))\right)$$

recursively until one can not find any triplet of distinct ex ante agents $h, i, j \in I$, such that:

$$\Delta(i,j) < \min\left(\Delta(i,h); \Delta(j,h)\right).$$

Provided this algorithm is processed, to any threshold of the similarity index $\bar{\Delta}$ corresponds an unambiguous partition π of the *ex ante* agents (this is obvious from the stop rule of the algorithm).

3 Data, estimation and results

3.1 Data

Our data are constituted of all European patents applications, one inventor of which at least declared an address in (metropolitan) France, and the priority date of which is between January 1977 and August 2003 included.[5] All non metropolitan French inventors

[4] We shall see in the next section how such a threshold can be chosen on the basis of the minimization of errors in a benchmark sample.

[5] These data are an extraction of the EP-INV database produced by CESPRI-Università Bocconi. For more details on the data see Lissoni et al. (2006).

of these patents are not considered. The dataset counts 122,157 patents and 236,824 inventor×patent occurrences. The latter set of observations represents our list of identities or *ex-ante* inventors, I. The variables used for computing similarity scores are presented in Table 1.

[Table 1, around here]

3.2 Homonymy

We are inclined to think that spelling errors are very limited in the French context. Therefore we will not consider here the similarity of inventors with different names and first names though our methodology also applies to such an issue. Therefore our empirical exercise will only deal with the homonymy problem. The inventor's name and first name is the alphanumerical variable X^1. Assuming no spelling error, then the probability that any agent changes her name or first name is zero ($\varepsilon^1 = 0$ and thus $\beta^1 \to \infty$).[6] In the mean time we still want to use the information of the relative frequencies of names and first names since they significantly influence the probability that i and j are in fact the same person. Therefore, we will compute :

$$\tilde{\Delta}(i,j) = \frac{1}{\beta^1}\Delta(i,j) = \Omega^1(i,j)\,\Pi_{k=2,...\bar{k}-1}\beta^k\Omega^k(i,j), \tag{4}$$

as similarity scores instead of $\Delta(i,j)$, and impose transitivity as exposed in Algorithm 1. Avoiding computing β^1 is fully consistent here, because in practice we will only compare the similarity of agents i and j who have same name and first name.

3.3 The benchmark

A list of French faculty members were matched with the patents on the basis of the name and first names of their inventors. Internet verification and phone calls were proceeded to these faculty members so as to enquire whether they really invented the patents in which an inventor with the same name and first name as theirs were mentioned. All in all, reliable information were collected on 445 French scholars.[7] Their positive and negative

[6] Assuming agents do not voluntaritly change their ID from time to time.

[7] We are indebt to the KEINS project and BETA at the University of Strasbourg for nicely letting us use these data.

declarations on whether they are or are not the inventors of these patents someone of the same name and first name invented are transformed into assertions whether some *ex ante* agent i and some other agent j who have the same name and first name are or are not the same person. Such process gives us $4,989$ matches, among which $4,567$ are validated matches ($i = j$) and 422 are incorrect matches ($i \neq j$). These matches and mismatches of homonyms can be used as a reliable benchmark in order to infer the quality of any assertion toward a matching of *ex ante* inventors. We can compute an error index, as the share of positive errors ϵ_1 (proportion of incorrectly predicted mismatches), the share of negative errors ϵ_2 (proportion of incorrectly predicted matches) as well as any linear combinations of these two shares

$$\phi(\theta) = \theta\epsilon_1 + (1-\theta)\epsilon_2, \tag{5}$$

with $\theta \in [0,1]$, which accounts for any weighting schemes of the willingness to avoid the two types of errors. A threshold on $\tilde{\Delta}(i,j)$ that would minimizes such value for any weighting scheme θ is noted $\bar{\Delta}(\theta)$.

3.4 Estimations

3.4.1 Initialization, recursive computations and convergence

Given the initial list of $236,824$ identities, we rely upon the name, the first name and the full address information to obtain an initial partition π^0 of identities.[8] It counts $127,605$ inventors. That initial aggregation allows us to compute initial conditional probabilities ε^k for each variable k (the frequency that, in two different identities, any agent keeps the same value of variable X^k, divided by the probability of the reverse). First values of the similarity scores $\tilde{\Delta}$ (also applying the transitivity algorithm) are then computed for the $898,682$ couples of homonyms.[9]

It should be noticed that, in this first round, the ε^k are underestimated since identities are not yet sufficiently aggregated in the "true" partition. Therefore we recursively process

[8] The full string reporting the city and the street address is considered. The probability that two persons with exactly same name and first name have the same address (i.e. they live in the same building) can be reasonably assumed to be equal to zero.

[9] Without relying on the location data at this stage because the address were used in defining the identities and so the probability to move is here null by assumption.

identities allowing to progressively determine the identities and the ε^k. This approach allows us to minimize abusive aggregations of agents. At each stage, we rely upon the benchmark to compute a threshold minimizing a "conservative" linear combination of errors, that is giving significantly more weight to negative errors, setting $\theta = 0.2$. In the initial round, an even more conservative θ ($\theta = 0$) is used to avoid negative errors given that we still have low confidence in the computations of ε^k, provided also that the information on the location of inventors has not yet been taken into account (see the previous footnote).

Thus at each stage t, the partition of *ex post* agents obtained in the previous period, π^t, is considered. New conditional probabilities ε^k, new similarity scores and new threshold $\bar{\Delta}(0.2)$ are computed. All pairs of agents the similarity score of which is above the threshold are aggregated within elements of π^{t+1}. That partition defines a new population of *ex post* agents for the next stage. This process is repeated until it converges to a partition π^* which will constitute the final set of *ex post* inventors.

3.4.2 Results

Table 2 displays the recursive computations of the conditional probabilities and the partitions, which convergence to a stable ε^k and a partition π^*.

[Table 2, around here]

The first computed partition π^1 counts $107,615$ agents. It is likely to not be sufficiently aggregated given the very conservative strategy adopted in the initial step ($\theta = 0$). However, the process proves to converge quickly since the next step partition π^2 already corresponds to the equilibrium one.

Table 3 reports the different sources of the variance of the similarity index. In order to compute those sources, there have been considered only values of $\tilde{\Delta}(i,j)$ unchanged by transitivity algorithm.[10] These values result of the actual similarity score calculated for each of the variables and are thus the only ones really informative. The values reported in the table show that the variance of the total score is affected mainly by the variance characterizing the city, X_3, and the applicant, X_2, variables. These two considered together

[10] About one quarter of the similaity indexes are imposed by the transitivity algorithm.

count for more than 40 per cent of the total similarity index variance. Then, the technological fields, X_4, and citation links, X_5, follow in the ranking. Therefore, city location is the most important information for defining identities in our sample.

[Table 3, around here]

Figure 1 reports the value of the error index as a function of the threshold, for various weights given to positive errors θ.[11] As expected, when $\theta = 0$, the error index is always increasing with the chosen threshold: the highest the threshold, the less the number of incorrectly predicted matches (negative errors). On the contrary, when $\theta = 1$, the error index is always decreasing with the threshold: the lower the threshold the less the number of incorrectly predicted mismatches (positive errors). For $0 < \theta < 1$, the error index decreases, reaches a minimum and then increases up to some limit. It appears that the value of the threshold $\bar{\Delta} \simeq \exp\{14.65\}$ corresponds to the minimal error index or a wide range of θ. When $\theta = .2$ (our preferred value), the error index is equal to .0189, that is the weighted average number of errors among possible ones is less than two percent.

[Figure 1, around here]

4 Conclusion

This paper proposes a Bayesian methodology to treat the who's who problem arising in individual level data sets such as patent data. The basic idea is to estimate the probability that two individuals are the same given some other observations. To do that, we rely upon a Bayesian approach providing a method for estimating the probability that two identities correspond to the same person. It relies on an estimation of the probabilities that *ex post* agents (e.g. inventors) change their observables (e.g. technological subfield) from one identity (e.g. invention occurrence) to the other. Doing so, we assign a similarity index to each couple of *ex ante* individuals and, given a threshold, we identify *ex-post* identities (i.e. the partition of the initial list of identities grouping together the ones which correspond to the same person).

[11] The equilibrium partition and the corresponding similarity indexes have been considered to draw the figure.

To our knowledge, it represents the first attempt to give a theoretical basis to the treatment of the who's who problem. In particular, the proposed methodology is able to overcome the main drawbacks of the approaches developed till now. First, our methodology takes in account the relative frequencies of each observation and of each variable. Doing so, we are able to fully exploit all the information contained in the data. Second, in computing the similarity index it relies on an endogenous procedures avoiding the exogenous and arbitrary scoring methods adopted by previous contributes.

Moreover, we assess the methodology developed referring to the set of all French inventors appearing on the EPO applications from 1978 to 2003. We define a recursive algorithm that permits to identify a stable partition of *ex-post* inventors and, using a benchmark dataset, provides also a measure of the weighted share of negative and positive errors. Our preferred specification allows us to leave a weighted average number of errors less than two percents of the potential ones.

References

Carayol, Nicolas, and Pascale Roux. 2008. "The Strategic Formation of Inter-individual Collaboration Networks. Evidence from Co-invention Patterns." *Annales d'Economie et de Statistique*, 89/90: 275-302.

Fleming, Lee, Santiago Mingo, and David Chen. 2007. "Collaborative Brokerage, Generative Creativity, and Creative Success." *Administrative Science Quarterly*, 52(3): 443-475.

Griliches, Zvi., ed. 1984. *R&D, Patents, and Productivity.* Chicago: University of Chicago Press.

Jackson, Matthew O., and Brian Rogers. 2007. "Meeting Strangers and Friends of Friends: How Random are Socially Generated Networks?" *American Economic Review*, 97(3): 890-915.

Jaffe, Adam, and Manual Trajtenberg, eds. 2002. Patents, *Citations and Innovations: A Window to Knowledge Economy.* Cambridge: MIT Press.

Lissoni, Francesco, Bulat Sanditov, and Gianluca Tarasconi. 2006. *The Keins Database on Academic Inventors: Methodology and Contents*, Cespri - Università Bocconi working paper 181.

Trajtenberg, Manuel, Gil Shiff and Ran Melamed, 2006. "The "Names Game": Harnessing Inventors' Patent Data for Economic Research." National Bureau of Economic Research Working Papers 12479.

Scherer, Frederic M. 1982. "Inter-industry Technology Flows and Productivity Growth." *The Review of Economics and Statistics*, 64(4): 627-634.

Schmookler, Jacob. 1966. *Invention and Economic Growth.* Cambridge: Harvard University Press.

Singh, Jasjit. 2005. "Collaborative Networks as Determinants of Knowledge Diffusion Patterns." *Management Science*, 51(5): 756-770.

5 Tables and Figures

Variables	# of modalities	Entropy
X_1 : same first name & name	$95,680$	0.95
X_2 : same assignee	$22,550$	0.72
X_3 : same city	$11,570$	0.72
X_4 : same IPC (6 digits)	$5,283$	0.85
X_5 : at least one citation link	$158,183$	0.93

Table 1. The variables used to build the similarity scores, the number of different groups of agents and normalized entropy indexes.

t	$\#\pi^t$	$\log\bar{\Delta}$	mean $\log\tilde{\Delta}$	std. $\log\tilde{\Delta}$	ε^1	ε^2	ε^3	ε^4	ε^5
0	127,605	17.63	14.30	4.64	-	0.2190	-	0.6671	0.8716
1	107.415	12.38	23.28	6.55	-	0.2757	0.1906	0.6976	0.8818
2	102,376	12.38	23.28	6.55	-	0.2757	0.1906	0.6976	0.8819
3	102,376				-				

Table 2. Convergence of the recursive computing of the *ex post* agents and the conditional probabilities.

	$\log \tilde{\Delta}$	X_1	X_2	X_3	X_4	X_5
Variances	52.07	1.30	10.86	12.25	6.88	3.19

Table 3. The sources of the variance of the similarity indexes (only value not imposed by transitivity algorithm are taken in account).

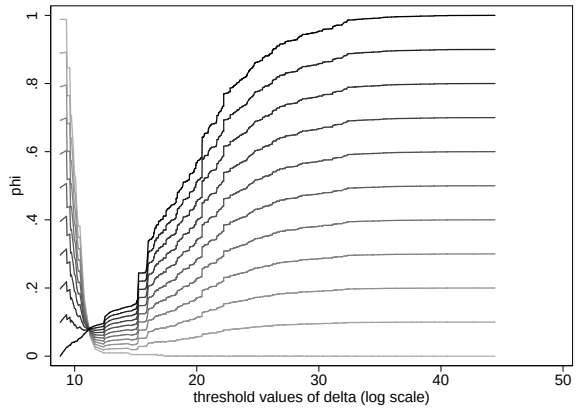

Figure 1. Computed values of the error index ϕ as the threshold of the similarity index $\bar{\Delta}$ increases, for different values of θ, 0.1 increments between its two bounds: $\theta = 0$ (black line, considers only negative errors, i.e. incorrectly predicted matches) and 1 (lightest grey line, considers only positive errors, i.e. incorrectly predicted mismatches).