The "Names Game": Harnessing Inventors Patent Data for Economic Research

By

Manuel Trajtenberg, Gil Shiff and Ran Melamed

Working Paper No. 10-2006

August, 2006

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

The "Names Game":
Harnessing Inventors Patent Data for Economic Research

Manuel Trajtenberg, Gil Shiff and Ran Melamed

The Eitan Berglas School of Economics
Tel Aviv University

August 10, 2006

Acknowledgements:
This project has benefited enormously from the work of a group of extremely talented and dedicated research assistants, primarily Michael Katz, Alon Eizenberg, and Ran Eilat. Useful comments were provided by participants in numerous seminars, particularly at the NBER. We gratefully acknowledge the financial support of the National Science Foundation grant SES-0527657, the Israeli Science Foundation Grant 1289/05, the Samuel Neaman Institute through its STE Program, and the Sapir Center.

Abstract

The goal of this paper is to lay out a methodology and corresponding computer algorithms, that allow us to extract the detailed data on inventors contained in patents, and harness it for economic research. Patent data has long been used in empirical research in economics, and yet the information on the identity (i.e. the names and location) of the patents' inventors has seldom been deployed in a large scale, primarily because of the *"who is who"* problem: the name of a given inventor may be spelled differently across her/his patents, and the exact same name may correspond to different inventors (i.e. the "John Smith" problem). Given that there are over 2 million patents with 2 inventors per patent on average, the *"who is who"* problem applies to over 4 million "records", which is obviously too large to tackle manually. We have thus developed an elaborate methodology and computerized procedure to address this problem in a comprehensive way. The end result is a list of 1.6 million unique inventors from all over the world, with detailed data on their patenting histories, their employers, co-inventors, etc. Forty percent of them have more than one patent, and 70,000 have more than 10 patents. We can trace those multiple inventors across time and space, and thus study the causes and consequences of their mobility across countries, regions, and employers. Given the increasing availability of large computerized data sets on individuals, there may be plenty of opportunities to deploy this methodology to other areas of economic research as well.

JEL: O30, C81, C88
Key words: Patents, inventors, mobility, computer software.

Manuel Trajtenberg
The Eitan Berglas School of Economics
Tel Aviv University
Tel Aviv, Israel
manuel@post.tau.ac.il

Gil Shiff
The Eitan Berglas School of Economics
Tel Aviv University
Tel Aviv, Israel
shiffgil@post.tau.ac.il

Ran Melamed
The Eitan Berglas School of Economics
Tel Aviv University
Tel Aviv, Israel
ranmelam@post.tau.ac.il

Table of Contents

I. Introduction

The main goal of this paper is to describe the way by which we can harness detailed data on inventors contained in patents so as to make them usable in economic research. We outline a methodology (and corresponding computer algorithms) for matching names that we have developed for this purpose, which may prove useful also in other contexts. Indeed, given the increasing availability of large computerized data sets on individuals, there may be plenty of opportunities to deploy this methodology in other areas of economic research as well.

Undoubtedly the richest source of data on worldwide innovation and technical change is patents, with millions of records offering detailed information of what was invented, by whom, in which fields, where, linkages to previous innovations, etc. Indeed, patents are one of the most extensive and detailed sources of data on ***any*** aspect of economic activity, and one that has far reaching implications for micro and macro economic performance. Mindful of its potential, economists tried already in the 1960s to deploy patent data in a large scale to the study of technical change. The computerization of the patent file in the early 1980s, the linking to Compustat, and the development of indicators based on patent citations in the 1990s constituted major advances in the ability to master such wealth of information for economic research. The project described here is a further step in that progression, namely, taking advantage of data on ***inventors*** appearing in patents in order to significantly expand the range of phenomena that can be investigated with the aid of these data.

The idea of using patent data in a large scale for economic research goes back to the seminal work of Schmookler (1966), followed by Scherer (1982), and Griliches (1984).[1] The work of Schmookler involved assigning patent counts to industries (by creating a concordance between patent subclasses and SICs), whereas Griliches' research program at the NBER entailed matching patents to Compustat firms. In both cases the

[1] This is not meant to be a survey but rather we just highlight wide-scale research projects that put forward distinctive methodologies of patent data construction, and had a significant impact on further research. For a survey of research using patent data, see Griliches (1990).

only data item used, aside from the match itself, was the timing of the patent (i.e. the grant or application year), such that in the end the patent data available for research consisted of patent counts by industries or firms, by year. Of course, it is the ***linking out*** of such data that made it more valuable, since it could then be related to the wealth of information available on the industries/firms themselves. The project that Scherer undertook involved classifying a sample of 15,000 patents into industry of origin and industries of use, by the textual examination of each patent. The result was a detailed technology flow matrix, that again could be linked out to external data, such as R&D expenditures on the one hand, and productivity growth on the other hand.

One of the major limitations of these and related research programs, extremely valuable as they had been, was that they relied exclusively on simple patent counts as indicators of innovative output. However, it has long been recognized that innovations vary enormously in their technological and economic "importance", "significance" or "value", and moreover, that the distribution of such "values" is extremely skewed. The line of research initiated by Schankerman and Pakes (1986) using patent renewal data clearly revealed these features of the patent data (see also Pakes and Simpson, 1991). Thus, simple patent counts were seriously and inherently limited in the extent to which they could faithfully capture and summarize the underlying heterogeneity (see Griliches, Hall and Pakes, 1987). A further (related) drawback was of course that these projects did not make use of any of the other data items contained in the patents themselves, and could not do so, given the stringent limitations on data availability at the time.

Keenly aware of the need to overcome those limitations and of the intriguing possibilities opened by patent citations (as revealed for example in Trajtenberg, 1990), Rebecca Henderson, Adam Jaffe and Manuel Trajtenberg undertook work aimed at demonstrating the potential usefulness of citations for a variety of purposes, primarily as indicators of spillovers (Jaffe, Trajtenberg and Henderson, 1993), and as ingredients in the construction of measures for key features of innovations such as "importance", "originality" and "generality" (Trajtenberg, Jaffe and Henderson, 1997). They used for these projects relatively small samples of patent data that were acquired and constructed

with a single, specific purpose in mind. However, as the data requirements grew it became clear that it was extremely inefficient, if not impossible, to carry out a large-scale research agenda on such a piece-wise basis.

Joined by Bronwyn Hall, Jaffe and Trajtenberg undertook to construct a comprehensive patent data file comprising detailed information on each patent as well as a series of indicators based on citations, that could not only account for (at least some of) the heterogeneity of patents, but also allow us to link patents over time and space. The result was the so-called "NBER Patent and Citations Data", which has been opened for general use since 2001 (see http://www.nber.org/patents/). The data comprise detailed information on almost 3 million US patents granted between January 1963 and December 1999, all patent citations made between 1975 and 1999 (over 16 million), and a reasonably broad match of patents to Compustat (the data set of all firms traded in the US stock market). A book followed soon after (Jaffe and Trajtenberg, 2002), containing many of the authors' previous articles on patents, as well as a CD with the complete data. The availability of these data has greatly stimulated research in this and related areas, and there are by now scores of papers and ongoing projects using it.

However, an important piece of information appearing in patents has not been used often in research so far, still less on a major scale, and that is the identity of the inventors themselves. As can be seen in Appendix 1, the front page of a patent contains the names and locations of each of the inventors that took part in that invention (on average there are 2 inventors per patent). The locations refer to the private domiciles of the inventors, not the address of the assignee (on the latter there is separate information).

If we could unequivocally identify each inventor (e.g. if each had an ID number), then we could follow the patenting history of each of them, trace their mobility, etc. Suppose for example that the inventor John Fields is issued a number of patents over his active life; for each patent we have his address at the time, the firm (if any) for which he worked (and hence the legal entity to which the patent was assigned), the identity of his co-inventors, and the rest of the information on the patent itself, as it appears in the

NBER Patent and Citations File. Observing the addresses of John Fields appearing in two consecutive patents, we could establish whether he moved or not sometime between the application dates of the first and second patents. He may have stayed put, or he may have moved counties, cities, states, and even countries. By the same token we could observe whether John Fields moved assignees, changed technological areas, and worked with different teams of co-inventors.

We also observe various indicators of the "quality" of each of his patents (such as citations received, generality, originality and number of claims), and could follow those indicators over time. Thus for example, preceding and following each decision node of "move/stay put" in terms of assignees and/or geographic location, we know what John Fields "innovational capital" is, as well as that of his partners. There is a wide array of interesting research questions that could be addressed if such data were available:

- Study spillovers by tracking the movement of inventors across countries, regions, assignees, type of institutions, and technology fields.
- Which inventors tend to move, in each of these dimensions? E.g. do "better" inventors tend to move more frequently (perhaps in order to achieve a better match), or the other way around?
- How does moving impact the future productivity of those inventors? That is, are the innovations that inventors make after moving more "valuable"? If so, what is the mechanism – better sorting? Does being exposed to a new/different environment result in new/better ideas?
- How does the mobility of inventors impinge on the innovative output of their employers? Which firms tend to lose inventors, which ones tend to gain? Is the net gain or loss what counts, or rather the turnover?
- How do teams influence mobility, and the subsequent productivity of inventors? Do different types of firms encourage different patterns of collaboration, which in turn may affect their own research productivity? Can we track the formation and evolution of "social networks" of inventors, and their impact?

- How do the above patterns vary across countries and over time? Is "brain drain" vs. "brain gain" really the issue, or is it rather the ability of regions to serve as "hubs" for inventors to come and go, and generate spillovers in the process?
- How do innovative clusters such as Silicon Valley emerge? Where do the inventors that form this type of clusters typically come from? Are they mostly first-time inventors?
- What are the policy implications of all these phenomena?

These are some of the issues that could be addressed, but surely there are many more. The research opportunities opened up by harnessing the inventors' data are undoubtedly far reaching and exciting, yet there have been very few attempts to do so on a large scale (see Table I.1 below), with good reason: a major stumbling block is that we cannot identify from the data as is *"who is who"* among the inventors, due to two fundamental problems. First, the name of the same inventor may be spelled slightly differently across some of his/her patents, it may come with or without the middle name and/or the initial, with or without surname modifies, etc. Thus, a name such as Tra***j***tenberg may be spelled in one patent with a "***j***", in another with a "***ch***" (i.e. Tra***ch***tenberg), and likewise for "Manuel" and "Emanuel". Secondly, suppose that the inventor name in one patent is exactly the same as the inventor name in another patent – do the two correspond necessarily to the same person? We don't know, and cannot infer it just from the name: this is the "John Smith" problem, that is, different inventors having exactly the same name may appear in various patents, and we need to be able to tell them apart.

Absent a way of dealing systematically with these issues the data on inventors is essentially useless, since whatever the shortcut strategy that one may adopt (e.g. match any two patents with exactly the same name, ignore all spelling variations, etc.), it would be riddled with error, and moreover, it would be impossible to assess the true extent and nature of those errors. Tackling these problems properly (and in finite time) is extremely difficult, for two reasons: first, the sheer size of the data, which consist of over 4 million

"records";[2] second, almost half of the inventors are located outside the USA, and foreign names, particularly East-Asian ones, present idiosyncratic problems of their own which require careful treatment. It is therefore clear that any attempt to address the *"who is who"* problem must rely on automated, computerized algorithms, and that there are significant economies of scale in doing so.

Aided by a very talented and dedicated team of research assistants,[3] Trajtenberg undertook back in 2002 to develop a "computerized matching procedure" (CMP) that would tackle head on the *"who is who"* problem, and render a list of unique inventors. Joined later by Shiff and Melamed,[4] and after 4 years of intensive efforts, the project has reached fruition: we have arrived at a well-performing and reasonably accurate CMP, produced a list of unique inventors, attached to it detailed data on the inventors' patenting histories, and probed the use of the data by conducting preliminary studies of inventors' mobility. These data will soon be opened to all researchers, hopefully encouraging a new wave of studies addressing the sort of research questions posed above.

Over the past 3-4 years there have been a significant number of research projects attempting to take advantage of inventors' data, most of them using relatively small samples, and thus being able to do the matching with the aid of ad hoc, manual methods. There have also been a few attempts to use large scale inventors' data, having to develop for that purpose some sort of computerized procedure. Table I.1 summarizes this emerging literature: Singh (2003) tackles the *"who is who"* problem head on, using as matching criteria the same (identical) first and last name, middle initial, and patent subcategory. Jones (2005) relies just on the names (again, first, last and middle initial),[5] Kim, Lee and Marschke (2005) use a variant of Trajtenberg (2004) but without numerical scoring, and Fleming and Marx (2006) rely on the frequencies of last names and the

[2] Each record is a unique combination of a patent and an inventor. Recalling that the NBER data contains over 2 million patents, and that each has on average 2 inventors, the multiplication gives the number of records in the Inventors file.
[3] They included Michael Katz, who did most of the Benchmark Israeli Inventors Set (see Section VI), Alon Eizenberg, who developed the "Mark I" CMP, and Ran Eilat, who developed parts of the final version of the CMP.
[4] Gil Shiff and Ran Melamed are currently graduate students at the Eitan Berglas School of Economics.
[5] Jones then matches a (reduced) list of US inventors with external data sources to obtain their ages.

overlap of co-inventors.[6] These projects have greatly increased our understanding of the potentialities of the inventors' data, shedding light in so doing on interesting aspects of inventors' mobility and related issues. Thus, they should be regarded as important stepping stones towards the development of a more comprehensive and accurate matching methodology, as the one attempted here.

Table I.1
Papers Using Patent Inventors Data

#	Authors	Data Source	Focus of research	Matching algorithm	# of inventors
	Large-scale patent data				
1.	Singh (2003)	NBER Patent file, USPTO, 1975-2002	Mobility of inventors, diffusion and social networks	Same 1st & last names, middle initial, patent sub-category (2 digit)	1.7 million
2.	Kim, Lee & Marschke (2005)	USPTO, NBER Patent File, etc. 1975-2002	Mobility from Universities to Industry	Similar to Trajtenberg (2004), w/t scoring	2.3 million (thru 2002)
3.	Jones (2005)	NBER Patent file, 1963-1999	Changing "burden of knowledge" of inventors; team work	Identical 1st & last names, and middle initials	1.4 million
4.	Fleming & Marx (2006)	Extended NBER Patent File, thru 2002	Employment changes of US inventors, non-compete agreements	Frequencies of names, + overlaps of co-inventors	2 million (thru 2002)
	Smaller samples				
5.	Stolpe (2001)	1,398 US patents, 1975-95	Mobility of inventors and spillovers in LCD technology	Acknowledges problem of lack of algorithm.	2,116
6.	Rosenkopf & Almeida (2003)	Patents of 74 semiconductor firms, 1990-95	Firm alliances and the mobility of inventors	NA	NA
7.	Song, Almeida and Wu (2003)	Patents of semiconductor firms, 1975-99	Learning by hiring, move of inventors from US to non-US firms	Exact names matched, plus manual/heuristic checks	180
8.	Crespi, Geuna & Nesta (2005)	PatVal, EPO, 1993-1997	Mobility of academic inventors	Survey	9,000
9.	Hoisl (2006)	Survey German inventors. Pat Val, 1977-2002	Mobility and productivity of inventors	NA	3,049 / several hundred

[6] Over the past 3-4 years Trajtenberg presented in numerous seminars the main thrust of the methodology, as well as first-cut results on inventors' mobility. Although he did not communicate the initial phases of the project via (quotable) working papers, the power-point presentations used in these seminars were made widely available and contribute to disseminate the methodological approach.

10.	Zucker & Darby (2006)	USPTO, 1976-2004	Careers of star scientists	Names, CVs	1,838
11.	Agrawal, Cockburn, & McHale (2003)	USPTO, NBER Patent file, 1990-2002	Social capital effect on knowledge spillovers	Exact name for finding self-citations	59,734 observations on movers
12.	Breschi &Lissoni (2003)	Italian inventors, EPO 1978-1999	Localized knowledge spillovers controlled by inventors network	Exact name, technological field	30,170
13.	Alcacer & Gittelman (2004)	Sample from USPTO, 2001-2003	The role of inventors and examiners in the generation of patent citations	Exact name, assignee, location	40,797

To sum up, this paper describes in detail the computerized matching procedure developed in order to tackle the "who is who" problem, dwells on a wide range of data issues regarding inventors' names as well as auxiliary data fields, discusses some key phases along the development process that may shed light on the quality and limitations of the CMP, and concludes with same basic statistics on the end product, i.e. the list of unique inventors. The intention is to provide extensive information on the matching method, in order to both allow prospective users of the data to assess its strengths and weaknesses, and to encourage further improvements in the CMP.

Longitudinal data on individuals have long been available from specially designed surveys; however, there are vast new opportunities to do research that focuses on individual data on a large scale (from administrative as well as commercial sources), which have been made available by recent advances in information and communications technologies (ICT). Tapping into those new sources often requires tracing individuals that are only partially identified in the data, and we hope that the methods presented here will prove useful in those other contexts as well. Of course, the deployment of powerful ICTs has been one of the main forces pushing scientific progress for decades, as best exemplified by the Genome project (and more recently by Proteomics). Economics has still a long way to go in embracing the possibilities opened by those fast changing technologies – this paper constitutes a small additional step in that direction.

II. Overview of the data and methodology

II.1 The data inputs

The raw data used in this project come from the NBER Patents and Citations Data File (see Hall, Jaffe and Trajtenberg, 2001), and in particular from the PAT63_99 file, the Inventor file and the CITE75_99 file.[7] PAT63_99 contains the main data fields from the front page of utility patents issued by the USPTO between 1963 and the end of 1999, as well as additional variables constructed with the aid of citations. The Inventors file consists of all patent-inventor pairs: patents typically have more than one inventor (the mean is 2), and hence each patent generates a number of records equal to the number of inventors appearing in it. The data fields in the Inventors file include the patent number, the name of the inventor and her address, as shown in Table II.1.

Table II.1
Data Fields in the Inventors File

I. Name of inventor
• Last name
• First name
• Middle name or initial
• Surname modifier (e.g. Jr., Sr., III)
II. Address of inventor
• Street address *(relevant only for unassigned patents or patents assigned to individuals)*
• City
• State *(US only)*
• Zip code *(only in some US patents)*
• Country

We merged the data of the Inventors file with the PAT63_99 file, thus creating a data set in which each record contains the information described in Table II.1 plus some of the key variables of the patent itself, such as the Assignee and Patent Class. Since as said each patent has on average about 2 inventors, the 2,139,313 patents in PAT63_99 for

[7] The methodology presented here can be applied of course to any set of USPTO patents; in fact, we intend to deploy it next to updated patent data running up to 2005.

1975-1999 generated 4,298,457 records in the new Inventors file;[8] this file constitutes the starting point of our work here.

II.2 A roadmap to the methodology

We sketch here the 2-stage matching methodology, and in later sections we dwell on the conceptual and technical details of each stage. As already mentioned, establishing "*who is who*" poses two fundamental problems: first, the name of the same inventor may be spelled slightly differently across her patents, and second, even if the inventor name in one patent is exactly the same as the name in another patent we don't know whether or not such name refers to the same person.

In order to address the first problem, namely the fact that the name of the same inventor may be spelled slightly differently from patent to patent, we adopt a two-track approach. The first is to "clean up" and standardize the names as much as possible, the second is to rely upon the "Soundex" system. The latter is a coding method adopted by the US Census in the 1930's, in order to tackle the problems posed by variations in the spelling of names (in particular of foreign names), for the purpose of indexing the census data. In our context the Soundex method offers a handy tool to group together all records that may potentially refer to the same inventor.

The second problem of determining who among the potential "suspects" displaying the same name (or equivalent names according to Soundex) refer in fact to the same person proved to be much more difficult. For that purpose we rely on pair-wise comparisons between any two "suspects", of a series of variables such as the middle name, the geographic location (e.g. zip codes, cities, etc.), the technological area (i.e. patent class), the assignee, the identity of the co-inventors, etc. If a data item is the same in two suspect records (e.g. if two records display the same address, or are in the same patent class, or share the same partners), then the pair is assigned a certain score. If the sum of these scores is above a predetermined threshold, the two records are "matched",

[8] The "gross" total was of 4,301,229 records. However, 2,772 records with missing last names or "duplicate records" were eliminated, rendering a net of 4,298,457 records. By duplicate records we mean records that have the same patent number and exactly the same inventor name, and hence are almost certainly mistakes.

that is, they are regarded as being the same inventor. Once that is done for all the pairs in the comparison set we impose transitivity, that is, if record ***A*** is matched to record ***B,*** and ***B*** to ***C,*** then the three are regarded as the same inventor (for a first glimpse of how the matching procedure works, see the 3 examples in Appendix 6).

We made some of the scores depend upon the "size" of the category used in the comparisons (such as city, assignee or patent class), and upon the frequency of names. Thus for example, if two suspects are located in the same city but the city is large they would receive a ***lower*** score on that account than if the two reside in a small town. The reason is simply that co-location in say New York is deemed to be less informative of the identity of individuals than co-location is small localities. In other words, the probability that two records displaying the same inventor name refer to the same individual is deemed higher if the two are located in a small town rather than a large one, and similarly for employers (i.e. assignees) and patent classes. The other parameter affecting the scoring system is the frequency of the names themselves: if a name is "rare" in terms of the number of times it appears in the Inventors file (e.g. Griliches versus Smith), then the score would be higher. The obvious reason is that two records displaying an identical "rare" name and appearing say in the same city are significantly more likely to refer to the same inventor, than if the name were a common one. The two criteria thus render a scoring ***matrix*** that relies on the size of cities, assignees, and patent class, and on the relative frequency of the inventor's name.

A serious issue that arose early-on was the absence of a clear benchmark against which to asses the performance of the proposed methodology: how could we know how well we were doing in matching names? How could we fine-tune it? In order to establish such benchmark, we constructed "manually" what we regard as a comprehensive set of unique Israeli inventors (i.e. inventors appearing in US patents that listed their addresses in Israel at least once). We could do that since their number was manageable (the initial list of "suspect" Israeli inventors consisted of about 15,000 records), we were familiar with the variations and frequencies of Israeli names, assignees and cities, and in those cases where the information was inconclusive we could resort to other sources of data,

and even contact the inventors themselves. The end result is a set of about 6,000 Israeli inventors, which surely is not entirely error free, but nevertheless can be regarded as sufficiently comprehensive and accurate to serve its purpose as a benchmark. Having developed in parallel the computerized algorithms to do the same, we could then fine-tune the methodology by "calibrating" the computerized results for Israeli inventors to the benchmark. This procedure is explained in detail in Section VI.

III. Stage 1: Grouping similar names using Soundex

The first stage of the procedure consists of identifying and grouping together all names/records that are deemed to refer potentially to the same inventor, e.g. Ben Grosman*n*, Ben Grossman and Ben*n* Grossman; we call such groupings ***"p-sets"*** – ***p*** for "potential," that is, potentially the same inventor. The key problem of course is that the name of a given inventor may be spelled in slightly different ways across the various patents in which the inventor appears. That may be due to typos, transcription errors, abbreviations, errors introduced by intermediaries handling the patent applications (e.g. patent lawyers), different perceptions with regard to the "correct" way of spelling a name (particularly relevant for foreign or non-English names), or even deliberate variations for strategic reasons.

Two types of problems arise in this context: The first is technical in nature, and refers to the appearance of all sorts of non-letter characters and symbols in the names, such as apostrophes (e.g. O'Brian), the bar mark (e.g. Jean-Jacques), numbers, spaces in the middle of names, etc. The second refers to differences in the actual spelling of names, e.g. Grosmann vs. Grossmann. In order to tackle the first we undertook the following steps (these changes affected 214,844 records):

- Eliminated ***all*** non-letter characters and symbols from the names (i.e. last name, first name, middle name and surname modifier), including numbers, apostrophes, commas, bar marks, and periods following initials**,** as well as the following symbols: &, :, /, `, and ;.

- Eliminated all spaces within names.
- All names were rewritten in capital letters.

Thus for example, the inventor name Klaus-Wolf Von Eickstedt was transformed into KLAUSWOLF VONEICKSTEDT.

In order to tackle the remaining spelling variations we needed a set of rules to "standardize" names, such that say the names Grosmann and Grossman would be identically coded, and thus (if having the same first name as well) be considered as part of the same ***p-set.*** In the second stage it may turn out that these refer to different inventors, but the point is that we would never know if the two records are not brought together to begin with and considered for a potential match.

The procedure we use is based on the "Soundex" algorithm for name standardization, developed by the US Census in order to overcome spelling variations of surnames (see http://www.archives.gov/genealogy/census/soundex.html). This algorithm transforms names into alphanumeric codes as follows: the first character in the code is the first letter of the original last name (upper case), followed by a 3-digit number, each representing a letter (consonant) appearing in the name. The digits are generated according to the following procedure:

1. Go through each of the letters after the initial, giving them numerical values as indicated in Table III.1 (these are called "scoring characters").
2. Ignore any letter if it is not a scoring character. The same holds true for spaces and punctuation marks. In particular, this means that all vowels as well as the letters **H, Y** and **W** are ignored.
3. If the value of a scoring character is the same as the previous letter then ignore it. Thus, if two '**T**'s come together in the middle of a name they are treated exactly the same as a single '**T**' or a single '**D**'. If they are separated by another non-scoring character then the same score can follow in the code. For example, the name PETTIT is P330: the second 'T' is ignored but the third one is not since a non-scoring 'I' intervenes.

4. Keep working through the name until you have created a code of 4 characters maximum (i.e. the first letter, and 3 digits); if there are less than 4 characters then pad zeros at the end to complete the 4 characters.
5. Optionally one can ignore a possessive prefix such as ‘Von’ or ‘Des’ (we did not implement this option).

Table III.1
The Coding Soundex System

Score	Letters (upper & lower case)
1	B F P V
2	C G J K Q S X Z
3	D T
4	L
5	M N
6	R
none	Vowels, punctuation, H, W, and Y

In order to increase the accuracy of the code we deploy the same algorithm, but with 6 additional digits rather than 3, that is, we continue the process indicated above (coding successive, non-identical consonants in the name) up to 6 digits, rather than stopping at three. Thus for example, **Trajtenberg** would be coded **T623** using the standard 3-digit Soundex, but we expand it to **T623516** (and even so we have not coded the last “g”). An important additional departure from the original Soundex system is that we implement the same procedure also for the inventor's ***first*** name, e.g. **Trajtenberg Manuel** would be fully coded as **T623516 M420000**. However, we ignore in the coding any other component of the name, such as middle name, middle initial, surname modifier (such as Jr. or Sr.), etc.; some of these items will play a role later on, when comparing pairs within the ***p-set***. Here are further examples:

GROSMANN KLAUS	G625500 K420000
GROSSMAN KLAUS	
GROSSMANN KLAUS	

HAGIAWARA MASAO	H260000 M200000
HAGIWARA MASOU	
HAGIWARE MASAY	

After assigning to each of the 4.3 million names/records a Soundex code in this manner, we proceed to form mutually exclusive *p-sets*, that is, groupings of identical Soundex codes, of which there are 630,887. Potentially then there are as many distinct, unique inventors, but of course we would expect that the second stage will rule out many inclusions, that is, some of the names within a *p-set* that were coded the same will turn out to belong to different inventors (see example 3 in Appendix 6). The 630,887 figure of *p-sets* rendered by the first stage is then a lower bound for the number of distinct inventors, and constitute the raw data to be fed into the second stage. If we had relied instead just on identically spelled first and last names, the number of mutually exclusive *p-sets* would have been 1.2 million. The use of Soundex then helps us guard against what we shall refer to as ***"Type I error"***, which occurs if we under-match records, i.e. if we miss records that should be compared to establish whether or not they match, but instead we regard them from the start as different inventors. A high incidence of Type I error would render a list that contains too many inventors presumed different (or "unique"), and therefore would lead to findings indicating too little mobility, too few spillovers, etc. relative to the true extent of these phenomena.

How well does the Soundex algorithm do in terms of avoiding Type I errors? There is no obvious way of assessing systematically the incidence of this type of error - if there were then we could improve the method accordingly. The only specific shortcoming of the Soundex system that we have been able to actually detect in the data is the following: in some cases there is a spelling problem precisely in the first letter of the last name, and that is of course fatal for Soundex, since it takes the initial as given. Thus, for

example, we have found that to happen with some Hebrew names (of Israeli inventors), as with "Jacob" and "Yacob", "Tsiddon" and "Ziddon", etc. Beyond Israeli inventors though we are unable to assess the extent to which this problem occurs, if at all.

There are other potential sources of Type I error that one can think of, and that Soundex could not overcome, but once again it is not possible to pinpoint them in the data and assess their incidence. Here are some such possibilities: First, Type I errors would arise if a nickname rather the full first name appears in some of the patents of an inventor (e.g. "Bob" instead of Robert, or "Bill" instead of William), whereas the full name appears in the other patents. In those cases patents of the same inventor will be assigned from the start to different *p-sets* since the Soundex code for the first name would be different, and therefore will not be matched.[9] Second, there may be cases in which the first, middle or last names appear in the wrong data fields (e.g. Smith Robert instead of Robert Smith) in some patents, and in the right places in other patents of the same inventor.

Third, there may be legitimate changes in the name of an inventor over time, and in particular changes in the last name due to changes in marital status. Thus, suppose that at some point during her career an inventor gets married and changes her last name. Even if nothing else had changed she will appear in two different *p-sets* and therefore her patents would not be regarded as belonging to the same inventor. A similar problem may arise if an inventor emigrates and changes the name to a "local" version.[10] To repeat, there is no way of knowing what the incidence of these potential sources of error is; based just on causal observation our impression is that the remaining Type I errors are very rare overall, and hence that Soundex does a good job at inclusion, i.e. at bringing together names that should be considered as potential matches. However, this issue surely requires further research.

[9] Thus for example, the Soundex code for Robert is R163000 vs. Bob: B100000, Bill: B400000 vs. William: W450000, etc. A possible solution might to "standardize" the most common nicknames to the original given name, but one cannot be sure that the nickname is not the "real" name to begin with.

[10] In some contexts this issue may be all too important to ignore, e.g. if focusing on research questions that involve women inventors or immigrant inventors to countries where adaptation of names to the local language is common practice.

We now turn to ***"Type II"*** errors, that is, those incurred when we end up matching records that belong in fact to different inventors. This will lead, of course, to "too few" inventors, and therefore to spurious mobility, spillovers, etc. As already mentioned, this turned out to be the predominant concern throughout, and therefore most of the methodological apparatus that we develop below is meant to tackle it. Here we address a more specific issue, namely, how the use of Soundex impinges on the probability of incurring Type II errors, and what can be done to ameliorate it. In principle the second stage of the matching process (i.e. checking every pair of records within a given *p-set* to see if they refer to the same inventor) should take care of Type II errors, but it turns out that the Soundex method itself may induce Type II errors that would have not occurred otherwise. Here are some examples:

BROOK WILLIAM	B620000 W450000
BRYG WILLIAM	
BYERS WILLIAM	

GARCIA DAVID	G620000 D130000
GREIG DAVID	
GROSS DAVID	

Clearly, there is no way that Brook, Bryg and Byers refer to the same inventor, but they happen to have the same first name, and the three records turned out to have enough in common otherwise to have passed the tests of Stage 2 (even though this is a very low probability event), thus ending up as the same inventor. The same happened with Garcia, Greig and Gross. The fact that Soundex grouped them together expanded the *p-set* too much and, given that Stage 2 is not (and cannot be) full proof, caused the error. How do we guard against Type II errors at this initial stage? One way is to use, as already mentioned, a 6-digit numeric code (after the initial) rather than 3 digits as envisioned in the original Soundex. As shown in Table III.2, this might make a difference in a large number of cases.

Table III.2 **Distribution of Names According to the Number of Non-Zero Digits in the Soundex Code**		
# of digits in the Soundex code	**# of last names**	**# of first names**
0	103,490	102,220
1	756,001	837,828
2	1,531,152	1,884,079
3	1,177,833	1,119,766
4	491,332	273,333
5	174,530	69,968
6	64,119	11,263

Thus, if we had used just 3 digits, over 700,000 last name records (i.e. those with 4 or more digits), and over 350,000 first name records would have received a code that does not sufficiently discriminate between different original names. Consider for example the case of two inventors named BERGEMONT and BRUGGEMANN: if using a 3-digit Soundex code, the two would have received the same code, B625. By expanding to 6 digits, each receives a different code, the first B625530, and the second B625500 (although in this particular case 5 digits would have been sufficient to discriminate).

The Soundex algorithm was originally designed to handle only last names, so the use of Soundex for the first name as well caused in many cases over-expansion of the *p-sets* and thus turned into an additional source of Type II errors. Therefore, we narrowed the *p-sets* definition from all records with the same last and first Soundex code, to all records with the same last name Soundex code and the same first name Soundex code ***only if the first name code is at least 2 digits long***. Otherwise (i.e., if the Soundex code of the first name has just ***one*** digit) the first names should be ***exactly*** the same in order to be included in the same *p-set*. Another way of guarding against Type II errors due to the use of Soundex is to demand stringent matching criteria in Stage 2 for cases with short Soundex-coded first names, which is in fact what we do in Section V.

Another potentially serious problem is that the Soundex system has been designed for English names, and seems to perform pretty well also for German names,[11] but definitely not for East-Asian names, in particular not for Japanese names (which constitute about 25% of all names). The following case exemplifies the poor performance of Soundex due to the abundance of vowels and the letter ***H*** in East-Asian names – to recall, Soundex ignores vowels and H, and hence assigns the same code, C000000, to all the following quite different last names:[12]

CHO
CHOE
CHOI
CHA
CHOY
CHIOO
CHIOU
CHIU
CHAE

In view of these limitations of the Soundex system, we treated East-Asian names in a slightly different manner during the matching process, setting in fact different matching criteria for inventors from Japan, Korea, China, Taiwan, Hong Kong and Singapore (see Section V and Appendix 4).

One way to assess *ex post* the impact of using Soundex (that is, after implementing Stage 2 of the matching process) is to count the number of inventors that were matched in spite of having different last names (but of course that do have the same Soundex-coded name). As we shall see below, the end result of the matching process is a list of 1,632,532 unique inventors; out of these, 23,548 were affected by Soundex (1.5%), that is, originally each had two or more different last names or different first names, and they would have not been matched together had it not been for the use of Soundex. The

[11] There are many cases of German names where the difference in spelling between two otherwise identical names is the presence or absence of an "E" (particularly following a "U"), such as in Mueller and Muller; Soundex solves easily this problem, since it ignores vowels altogether.

[12] Of course, different ***first*** names discriminate between some of the equally coded last names, but nevertheless many Type II errors are incurred. Tetsuo Wada of the Faculty of Economics at Gakushuin University, Japan, has long been working on the problem of matching Japanese inventors names, and may be able to contribute to further improving in that regard the CMP developed here.

distribution of different initial names per inventor is shown in Table III.3.

Table III.3
Distribution of Number of Different Names per Inventor

Number of initially different names	Number of inventors
1	1,608,984
2	21,994
3	1,259
4	164
5	54
6	28
7	16
8	9
9	6
10	5
11+	13[13]
Total	**1,632,532**

To clarify, the overwhelming majority of unique inventors had ***exactly*** the same name to begin with, and hence Soundex was not relevant for them. However, 21,944 inventors had two different names, such that if Soundex would have not been used we would have not regarded them as the same inventor, and so forth for those with more than two original names.

In view of these figures one might conclude that the impact of Soundex is marginal (in terms of percentages it is indeed very small), and therefore perhaps not worth the trouble. A closer examination of the results indicates otherwise: the inventors affected by Soundex have by definition more than one patent, and hence their number should be contrasted to the 648,673 that have 2 or more patents, and not to the total of 1.6 million. More importantly, these inventors are inherently the more interesting ones, since they are by definition more prolific, and have patenting careers that can be traced over

[13] All of these 13 inventors are Japanese, and most of them constitute quite likely Type II errors, i.e. they were matched together but should not have been. In most cases the matching was done on the basis of the same city (see Appendix 3), and/or same Soundex-coded partners' names. Among those 13 inventors, 12 have 11-15 different names and one has 36 different names!

time and space. Ignoring them (that is, regarding their different patents as belonging to different inventors) would seriously affect our ability to study the type of phenomena that this whole enterprise is supposes to allow for. Furthermore, there is a big difference in this respect between the *ex ante* and the *ex post*: there was no way of knowing *ex ante* how much of a difference Soundex would make, and not using it (or some other coding system) would have left us in the dark regarding the incidence of Type I errors. Clearly, the better we do in terms of "cleaning" the names before using Soundex, the less Soundex will matter, but again, we may know what that implies only on hindsight.

To recap, Stage 1 consists of transforming the raw file of 4.3 million records into mutually exclusive *p-sets*, that is, groupings of records that have sufficiently similar names to be regarded as being potentially the same inventor. In so doing, we first clean-up and standardize the names (last and first names), and apply the 6-digit Soundex coding method to both the first and the last name of each record. Records with the same such alphanumeric code are grouped together into *p-sets*, for consideration in the second stage.

IV. Stage 2.a: Comparing names

Having grouped the standardized inventors' names in the first stage, the question now is how to decide whether or not each pair of records within a given *p-set* refers to the same inventor. As already mentioned, this turned out to be by far the most difficult task, and absorbed accordingly most of our efforts. The ensuing procedure involves comparing each pair of records within a given *p-set* according to a set of criteria, assigning a score to the dichotomous result of each comparison, and then setting decision rules based on the total score. We rely for that purpose on detailed information contained in each patent, data on each inventor and her co-inventors, and ancillary information derived from the patent file, such as the frequency of names, the size distribution of cities, firms, etc.

Just to get a sense of the scope of the problem of establishing *"who is who"* in the second stage, take as an example the proverbial John Smith case: there are 552 records with the Soundex code S530000 J500000, and a priori we have no idea how many

different inventors are included within this set. Using middle names narrows down the problem, and yet there are for example 78 records with the S530000 J500000 code and the middle name initial "W".[14] Again, there is no way of knowing *"who is who"* within that restricted set, unless one undertakes to develop a comprehensive, computerized system for comparing look-alike records.

IV.1 Auxiliary tools: name frequencies and category size

Before presenting the matching criteria, we dwell on two issues that affect the informativeness of the criteria used: name frequencies, and the size of the categories involved (for cities, assignees, and patent classes). Both family names and first names vary a great deal in terms of their observed frequency in the relevant populations, some being very common (e.g. Smith), others relatively rare. Clearly, such information is potentially very useful in deciding whether or not two records with the same Soundex-coded name refer to the same inventor: if the names are "rare" there is a priori a higher likelihood that they are indeed the same inventor, whereas the converse is true if the name is rather common. Implementing this idea would require building a measure of the frequency of the names appearing in our data, ***within their relevant populations***.[15] The problem is that our data comprises names of inventors from 165 countries, and hence doing that would require a massive effort that is well beyond the scope of the present project.

Short of using the (true) frequencies of each name within its population, we could compute the frequencies of names in our data, and use these as proxies. Surely such measures are informative, particularly for countries heavily represented in the data such as the US, Japan and Germany. However, for other countries not only the samples get much smaller, but there is no clear relationship between the size (population) of the country and the number of patents taken in the US. Thus for example, there are few patents from China or India, and hence the observed frequency of Chinese or Indian

[14] In the course of this project we have unwittingly learned a large number of "trivia", e.g. that the most prevalent middle name initial for John Smith is "W"…

[15] Thus for example, Fleming and Marx (2006) use name frequencies from the US Census website for the names of US inventors.

names in our data cannot be taken to be representative of how common these names are in their respective populations. For that reason we decided not to use the observed frequency of names in our data as a matching criterion of its own, but rather as an auxiliary tool determining the "strength" (i.e. the informativeness) of various criteria, including assignee, city and patent class. Take for example two records having the same Soundex-coded name that share also the same assignee: we shall regard this matching criterion (i.e. same assignee) to be more informative of a possible match if the name of the inventor is relatively "rare" than if the name is "common." We compute for that purpose the frequency of (Soundex-coded) names in the population of inventors in our data, and set a cutoff value: a name is considered "rare" if it appears at most 16 times, otherwise it is considered "frequent" (see Appendix 3).

We rely also upon the ***size*** of the assignee, the city and the patent class as an auxiliary criterion in setting the matching scores: belonging to a small entity is regarded to be more informative, and hence confers a higher score than belonging to a large entity, where size is taken to mean here the number of patents. The logic is simply that the probability that two different inventors will have the same name is higher in large entities than in small entities: two John Fields in IBM are less likely to refer to the same inventor than two records with such name in a small startup. The same goes for New York versus Boca Raton, and for a large patent class versus a small one. That is, two records exhibiting the same name are more likely to refer to the same inventor if the location is a small town, and/or the patent class encompasses a relatively narrow technological area.

We determine the size of cities, assignees, and patent classes according to the number of patents that each received in our data. This is not a self-evident choice: in principle one could use external data such as population for cities, or the number of employees for assignees (for patent classes there is no obvious size measure other than the number of patents). The reason we resorted to the number of patents is simply expedience: it would have been very hard to collect the necessary outside information, and link it up correctly with our data. In fact, this would have been virtually impossibly for the majority of assignees, in light of the difficulties encountered in matching their

names to Compustat (see Hall *et al* 2001). Besides, it is not clear which is the most appropriate (external) size measure – is it say the city's entire population, or just the "relevant" population of patenting inventors?

We thus compute the patent frequencies of each city, assignee and patent class, and set for each of them a cutoff level such that being below it makes the respective category "small" (and hence as more informative), and conversely a category is deemed "large" and hence less informative if it is above the cutoff value. In order to determine the cutoff values we examined closely in each case the distribution of patent frequencies, particularly around the mean and the median to see if these could serve as reasonable values (see Appendix 3 for a detailed discussion). Table IV.1 presents the median and the actual cutoff values for each matching criteria.

Table IV.1
Cutoff Values
(in terms of number of patents)

Category	**Median**	**Actual Cutoff**
Soundex-coded Names	23	16
City	1,382	1,382
Assignee	1,540	500
Patent Class	18,861	18,861

IV.2 The matching criteria

We now lay out in detail the use of matching criteria, and discuss their relative informational strength. Once again, given that two records exhibit the same Soundex-coded name and hence belong to the same *p-set*, we compare them according to whether or not various criteria hold for them, e.g. whether or not they share the same address, the same middle name, the same assignee, and so forth. Whenever a criterion holds the pair is assigned a score, and then the sum of the scores is compared to a threshold. Table IV.2 lists the data fields used in developing the comparison criteria.

Table IV.2 **List of Data Fields Used in Matching Criteria**
1. Name of inventor (*in addition to first and last)*:
Middle name *(name or initial)*
Surname modifier *(Jr., Sr., III, etc.)*
2. Location of inventor:
Street address *(relevant only for unassigned patents or patents assigned to individuals)*
City
State *(US only)*
Zip code *(only in some US patents)*
Country
3. Assignee
4. Technological classification:
Patent class
5. Patent citations
6. Overlap of co-inventors

The first three criteria which will be detailed below are the "strongest" and stand on their own, whereas the following ones depend upon the frequency of names and the size of the categories:

1. ***Full Address***

This criterion is met whenever two records share the same country, city and street address,[16] as in the following example:

record	patent	Last name	First name	Mid name	street	city	state
1	4211224	Kubach	John	S	1406 Milan RD	Sandusky	OH
2	4287794	Kubach	John	S	1406 Milan RD	Sandusky	OH
3	5404787	Kubach	John	S	1406 Milan RD	Sandusky	OH

[16] For U.S. addresses the Zip code can be used as well.

We consider this to be a very "strong" criterion, since it is extremely unlikely that two different inventors with the same Soundex-coded name reside in exactly the same address. One could conceive of cases in which that might be so, e.g. if by sheer coincidence two inventors with the same name live in the same apartment building, or if a parent and his/her son/daughter reside in the same house, happen to have the same Soundex-coded name, and the parent appears as inventor in one patent and the son/daughter in another. However, it is fair to assume that cases like these are extremely rare,[17] and therefore we view the ***full address criterion*** as "near-certain". Unfortunately, only about 11% of the records have a non-missing value in the street address field.

2. ***Self Citation***

Consider two patents, 1 and 2, sharing the same Soundex-coded inventor's name; the self-citation criterion is satisfied when patent 2, where Joe Doe appears as one of the inventors, cites patent 1, where the same Soundex code appears. Since the probability of self-citation is known to be significantly higher *ceteris paribus* than the probability of citing someone else's patent, then the converse must also be true, that is, if we observe a self citation then the two Soundex-equivalent names are likely to refer to the same inventor. In other words, a citation relationship, conditional upon the name of the inventor being the same in both records, significantly raises the probability that the two are in fact the same person.[18]

3. ***Shared Partners***

This criterion refers to the fact that collaborations among inventors are very likely to be persistent: if two patents share the same Soundex-coded name and the same co-inventor(s) Soundex-coded name, then the two quite probably refer to the same inventor. One may ask the question the other way around: suppose that two inventor names, Joe

[17] We know though of a few hundred records whereby family members are listed as inventors in the same patent, and in some cases reside in the same address. In order to avoid confusion we deleted these records from the data set.

[18] One could further improve the procedure in this respect, by iterating on this criterion, that is, run the matching procedure once, then run it again but with the citing and cited names as identified by the first run, and so forth.

Doe and Mary Beth, appear both in patent 1 and in patent 2, and suppose that we know that Mary Beth is one and the same inventor in both, but we are not sure whether the name Joe Doe in patent 1 stands for the same inventor as Joe Doe in patent 2. It is quite clear that the probability that Mary Beth will team up with two different inventors that have the same Soundex-coded name is exceedingly low.

4. *Full middle name / middle name initial / surname modifier*

The premise here is that the degree of informativeness of names (regarding the *"who is who"* problem) follows the following order: last (family) name, first name, middle name, middle name initial, surname modifier. Thus, given that two records share the same Soundex-coded first and last name, we further look into whether or not the two share also the same middle name and so forth.

About half the records in our data contain a non-missing value for MIDNAM; the ***full middle name*** criterion is satisfied whenever two records share the same Soundex-coded middle name, and the code is longer than one character, that is, it refers to a true name and not just to an initial. We regard this criterion as fairly strong, less so than full address or shared partners, but more than other criteria listed below. An obvious advantage of this criterion is the fact that a person's middle name is (typically) permanent, unlike location, assignee or technical field. The backdrop is that the middle name need not be consistently specified, that is, it may appear in one patent of the inventor but not in another.

In many records we observe just the initial rather than the full middle name, and hence we may not be able to tell for example, whether John W. Fields and John William Fields refer to the same inventor. As already mentioned, the ***full middle name*** criterion would not be satisfied for such two records, since we ignore initials in that context. However, the middle name's initial is informative in and of itself, and should increase the likelihood of a match.[19] We define the variable INITIAL as containing the first character

[19] One could argue that people may have a fixed tendency to specify their middle name either as an initial or as the full name. In such a case, a "John William Fields" and a "John W. Fields" would be ***less*** likely to

of MIDNAM: whenever two records in a *p-set* share the same (non-missing) value of INITIAL this criterion holds. We make the score associated with this criterion depend also on the frequency of the last and first names involved.[20]

Lastly, the ***surname modifier*** criterion is satisfied whenever two records share the same non-missing MODIFNAM value (a typical value for the MODIFNAM variable is "Jr."). Only 88,587 records in our data have non-missing values for this field.

5. ***Assignee***

The "assignee" is the organization to which the patent is assigned at issue (or reassigned later on). The assignee may be the firm/corporation in which the inventor works (these are the majority of cases), a Government agency, a University or other such organizations. Missing values for assignee indicate that the patent was unassigned or assigned to an individual. Clearly, if two patents exhibiting the same Soundex-coded name exhibit also the same assignee, it is more likely that the two refer to the same inventor than if the assignees were different. As already mentioned, if the assignee appearing in both records is "small" the same assignee criterion confers a higher score than if the assignee is "large." Likewise, for a given assignee size rare names get a higher score than common names.[21]

6. ***City***

This criterion is satisfied whenever two records sharing the same Soundex-coded name share also the same (non-missing) city (for U.S. inventors the ZIP variable serves

---

be the same person than two people named simply "John Fields" with no middle name information at all. Here we took the stance that the coincidence of the initials is informative, and thus assigned this criterion a positive score.

[20] In view of the difficulties mentioned above, we decided not to preclude the matching of records bearing different middle names initials, primarily because this would have confounded the use of transitivity. Although this may induce in some cases Type II errors, we found that enabling such matches yields better results than preventing them.

[21] One may claim that by using same assignee or same city as criteria for a match we may be introducing a downward bias in the very phenomenon that we would like to study, namely the mobility of inventors across assignees or locations. On the other hand, such information is clearly relevant, and it would be wrong to ignore it. The issue then boils down to the weight given to these criteria: by making them depend upon size and name frequency, we are clearly making it harder to match by them.

the same function).[22] As with assignee, we distinguish between large and small cities, and further differentiate the score by the frequency of names: a rare name in a small city carries more informational weight than a common name in a large city. An alternative suggested by Agrawal, Cockburn, & McHale (2003) and by Fleming & Marx (2006) would be to rely on Metropolitan Statistical Areas (MSA), and use it as a weaker matching criterion. Another possible refinement would be measuring the distances between the locations of any pair of records in the same *p-set,* and having the score on that account vary continuously with distance. That is, we would assign a higher score if the two suspect records are located in nearby cities and a lower one otherwise.[23]

7. *Patent class*

This criterion pertains to the affinity between records in technology space, as indicated by the patent classification system: inventors are likely to work in the same or similar technological fields over time, and hence are likely to obtain patents classified in the same patent class. To put it differently, two records exhibiting the same Soundex-coded name are more likely to refer to the same inventor if the patent class in both is the same. Note however that patents may be closely related and yet not be classified in the same (main) patent class, and hence this criterion is rather weak even in what it pertains to capture.[24] As with the previous two criteria, belonging to smaller patent classes is deemed to be more informative than belonging to larger ones.

V. Stage 2.b: The matching process

The first stage of the matching process consists of comparing each pair of records within a given *p-set* according to each of the above criteria, and assigning a "score" whenever the criterion holds. The scores are meant to reflect the strength of each criterion, that is, the extent to which the comparison according to that variable is thought

[22] "Same city" means the same city name in the same country, and in the same state if in the US. Obviously, he city criterion is relevant only if the stronger full address criterion was not used (the full address includes the city name).

[23] This manner of scoring (continuous vs. discrete) may be implemented also in the context of other criteria, such as proximity between assignees in terms of SICs, or technological proximity between patent classes.

[24] One could think of further refining this criterion, by using also cross-classification and not only the main patent class, and perhaps also field of search. It is not clear whether the extra effort is worth its while in terms of increases in matching accuracy.

to be informative. Thus for example, if two records having the same Soundex-coded name have the same full address then we are as sure as one can be that the two refer to the same inventor, and hence the score on that account will be the highest (and in fact in most cases it will be sufficient for a match). On the other hand, sharing the same patent class is a rather weak indicator, and hence the score on that account will be low and size-dependent. Once all comparisons have been made we sum up and compute the total score. If it exceeds a certain threshold, the two records are said to correspond to the same inventor and a match is performed.

Clearly, any numerical scheme of scores and thresholds would be inherently arbitrary, since we would be assigning a ***cardinal*** measure to what is essentially only an ***ordinal*** relationship. In other words, we can rank with a reasonable degree of confidence the different criteria in terms of how informative they are for matching inventors, but we can hardly be very precise in terms of how much any one criterion is "stronger" than the next in line. Nevertheless, we want to impute (cardinal) scores for the following reasons: first, to be able to sum up individual scores and use the resulting total as the final criterion for matching; second, to use the total score for diagnostic purposes (in fine-tuning the method, and in characterizing the degree of similarity between matched records); and finally, in order to use the total scores further down the line as weights in econometric estimation.

Following a lengthy and cumbersome process of extensive experimentation with alternative scoring schemes and corresponding thresholds, we settled for the one presented below, which seems to perform fairly well. However, we should keep in mind that this is by no means a full-proof scheme, and that there is as said an unavoidable measure of arbitrariness in the use of any such procedure. Quite clearly, there is no inherent meaning to the numerical values of the scores, but only in conjunction with the thresholds. Thus for example, a score of 100 for a given criterion *vis a vis* a threshold of 120 just means that this criterion by itself is not enough to ensure a match, but is quite "close" to it, so that in conjunction with just another "weak" criterion it would suffice. We could have normalized the scores relative to the highest threshold (and set the latter to

1), so that the scores could be interpreted as being fractional to the max threshold. However, we followed a different convention during the trial and error process, and we decided to stick to it entirely for pragmatic reasons.[25]

V.1 Thresholds

Since the scores are meaningful only *vis a vis* the thresholds, we start with the latter. Rather than having a unique threshold we specify three different threshold levels, depending on various characteristics of the names themselves. The alternative would have been to treat these characteristics as matching criteria, add their scores to the criteria listed above, and compare the total to a unique threshold. Obviously one can construct the scoring scales so as to make the two procedures exactly equivalent, but then again "history matters": the experimentation process that we followed led us to the present scheme, and we saw no reason to tinker with the computer algorithm, given that it works well as is.

The thresholds differ according to the extent to which the last and first names are informative in and of themselves: whether or not the names are ***exactly*** the same (as opposed to being the same Soundex-coded), and what is their length in terms of Soundex characters. Thus, the threshold level is lower the more similar the names are to begin with, and the more non-zero Soundex characters they comprise – clearly, longer Soundex-codes are more informative, a fact that is particularly relevant for East-Asian names. Table V.1 presents the criteria used to set the thresholds and their respective numerical values.

[25] Setting the scores as percentages of the threshold would require setting *different* scores for each of the different thresholds, hence further increasing the programming complexity and the room for error.

Table V.1 Thresholds	
Informativeness of names and determination of thresholds	**Threshold values**
• ***Exactly*** same first name (or Soundex-coded first name has at least 5 non-zero digits) *and* exactly same last name (or Soundex-coded last name has at least 5 non-zero digits)	100
• ***Exactly*** same last name (but not exactly same first name) or • Soundex-coded last name has at least 2 non-zero digits (but less than 5)	120
• All other cases	180

V.2 The scoring scheme

We categorize the various criteria into four "groups" according to their relative strength in conveying information for the matching decision, and assign to each group a numerical score, which should be interpreted in terms of the specified threshold levels. As mentioned before, the scoring of the criteria related to city, assignee and patent class depends both upon the frequency of names and upon size (computed as the number of patents of each category), as shown in Table V.2.

Table V.2 Size and Frequency Dependent Scores				
	Cutoff levels		**Score**	
	"Rare" name *(freq > 17)*	**"common" name** *(freq ≤16)*	**Below cutoff**	**Above Cutoff**
City	2,500	1,382 *(median)*	100	80
Assignee	2,500	500	100	80
Patent Class	30,000	18,861 *(median)*	80	50

Thus, if two records/inventors are located in the same "small" city (in the sense that less than 2,500 patents originate there), and their name is "rare", then on that account the pair would receive a score of 100. If the same pair would have been located in a "large" city (i.e. with more than 2,500 patents), the score would have been 80. The following example may help visualize this scoring scheme:

City	**City Size** *(number of patents)*	**Scores according to name frequencies**	
		John Smith *("common")*	**Aharon Trajtenberg** *("rare")*
Sacramento	1,217	100	100
Memphis	2,097	80	100
Los Altos	5,968	80	80

Thus two records with "Aharon Trajtenberg", both in Memphis, give a score of 100 to the same city matching criterion, since the 2,097 patents of Memphis are below the cutoff of 2,500 for the combination of city size and rare names; however, if the name would have been John Smith the score would have been just 80 (i.e. less informative). Table V.3 shows the complete scoring scheme.

Table V.3
Scoring Scheme
(threshold levels: 100, 120, 180)

Group	**Criterion**	**Score**
1	Exact same address	120
1	Self citation	120
1	Shared partners (co-inventors)	120
2	Full middle name	100
2	Initial of middle name for "rare" names[26]	100
2	"Small" assignee / rare names	100

[26] To recall, the cutoff level for names is 16, i.e. if the frequency of a name in the data is less than 16 it regarded as "rare", and the converse for names that appear 16 or more times.

2	“Small” city (or Zip) / rare names	100
3	“Small” patent class / rare names	80
3	“Large” assignee / frequent names	80
3	“Large” city / frequent names	80
4	“Large” patent class / frequent names	50
4	Surname modifier	50
4	Initial of middle name for frequent names	50

Thus, any of the criteria in Group 1 is sufficient to ensure a match if the last name of the two records compared is ***exactly*** the same, or if the Soundex-coded last name has at least 2 non-zero characters, since in such cases the threshold is 120 and so is the score that Group 1 criteria get. On the other hand, if the names are not very informative to begin with and hence the threshold is 180, then no single criterion is enough, and in fact for weaker criteria it would take at least two of Group 4 and one of Group 3 to ensure a match.

To recap, the procedure entails comparing every pair of records within a given *p-set* according to the various matching criteria, so that whenever a criterion holds the pair receives the corresponding score according to the table above. Finally, we compute the total score and compare it to the appropriate threshold, which in turn depends upon the characteristics of the name. If the total score exceeds the specified threshold we regard the two as the same inventor, and assign her a uniquely defined ID.

V.3 Transitivity

Stage 2 of the matching procedure entails making $n(n-1)/2$ pair-wise comparisons within each *p-set*, where n is the number of Soundex-coded names in the *p-set*. Each such comparison renders a discrete decision of whether to match or not, but then we may be confronted with the following conundrum: supposed that there are 3 Soundex-coded names in the *p-set*, *A, B,* and *C*, and that the comparisons indicate that *A* and *B* match, *B* and *C* match, but *A* and *C* do not – whom should we regard as being the same inventor?

Logic dictates that we should impose transitivity, that is, if *A* and *B* refer to the same inventor, and so do *B* and *C*, then *A* should match *C* as well, and thus the three of them should be regarded as one and the same inventor. This is not a trivial decision and certainly not an innocent one, particularly if the *p-set* is large;[27] however, it seems that transitivity is the only plausible course of action in such situations, which would render a logically consistent procedure.

VI. The Benchmark Israeli Inventors Set (BIIS)

One of the key problems facing the development of a computerized matching procedure (CMP) is how to assess its performance: on the one hand the file is far too large to allow for good enough sampling/random manual checks (and even then it is not clear how to conduct such tests), and on the other hand there is no natural or readily available benchmarks against which to compare the results. This is particularly troublesome in view of the fact that the procedure entails by necessity the choice of a series of discretionary matching parameters (primarily the matching scores *vis a vis* the thresholds, and the cutoff values for the size and frequency dependent categories): a priori considerations (as much as common sense) may help set starting values, but how are we to fine-tune them in order to optimize the procedure?

Mindful of this prime concern, and also of the need to engage in a learning-by-doing process on a manageable scale, we decided to tackle at first just the set of Israeli inventors, defined henceforth as inventors that had at least one USPTO patent with an address in Israel. Given that there were relatively few of them (about 6,000), and in view of our intimate familiarity with the country and its High Tech sector (which is the source of the vast majority of Israeli patented innovations), we could hope to be able to pin them down with a high degree of accuracy in finite time. Thus, the expectation was that doing a "manual" matching of Israeli inventors would result in a reliable and comprehensive set

[27] Consider for example the case where n=5, and hence there are 10 pair-wise comparisons to make between A, B, C, D and E. In principle it could be that the only comparisons that get a "passing score" are the 4 sequential ones (i.e. A to B, B to C, C to D and D to E), whereas the 6 others do not. In such case transitivity means that the end result is that all five are deemed to be the same inventor, even though most pair-wise comparisons fail to detect sufficiently similarity between them.

of Israeli inventors that could serve as benchmark for the CMP, and at the same time allow us to gather a great deal of know-how about how to design such procedure.

The starting point was thus all patents in which one of the inventors had an address in Israel (there were 13,565 such records); we then took the names of those inventors, and extracted ***all*** the patents bearing also their names (obviously with addresses in other countries as well), which brought the total to 18,807 records: these can be regarded as all the patent records associated with Israeli inventors (we refer to it as the "all inclusive set"). The goal was then to create a list of ***unique*** Israeli inventors that could serve as said as a benchmark.[28]

We proceeded by developing a first-cut CMP following similar (but much coarser) principles as those outlined above, deployed it on the all-inclusive Israeli set, and examined carefully the ensuing list one by one (in alphabetical order). Suppose that 3 records were "matched" by this method: we observed then 3 rows of data, each with the data fields of each of the 3 patents presumed to belong to the same inventor, including the corresponding name in each case, address, assignee, etc. We then applied specific knowledge of names and spelling, assignees, locations, etc. as much as discretion and common sense in order to decide whether or not the match was justified. In case of remaining doubts we looked for further clues in the patents themselves, and in a few hundred recalcitrant cases sought additional external information, including phone calls to dozens of individuals and firms.

This tedious, time consuming procedure was made even harder by the fact that in some cases the initial alphabetical sorting of names did not necessarily bring together (that is, in close proximity) all the names that needed to be considered for a match: Yakoby and Jacoby for example would not appear next to each other on the spreadsheet, and hence if they referred to the same inventor we could easily miss them. Awareness of

[28] Note that in this case not all the records would end up as part of the final set: if for example we start with inventor ***A*** having a patent located in Israel, and we extract a patent with inventor ***A'*** (i.e. with a name similar or even identical to ***A***) but with an address in another country, then if the comparison rules out that the two are the same inventor, the record belonging to ***A'*** just gets discarded from the set.

this problem brought us to develop heuristic rules to seek additional matches, particularly for some letters/initials (such as J and Y).

The end result was a list of 6,023 unique Israeli inventors and all their patents, totaling 15,316 records, which we can safely regard as being as comprehensive and accurate as possible. "Accuracy" here means that there should be very few Type II errors left, that is, as far as we know we have not matched together inventors that are in fact different individuals. As to Type I errors, we may have missed records when forming the all-inclusive set, and as said there may still be cases such as "Yacoby and Jacoby" which we did not identify. Note however that since the key issue is the performance of Stage 2 of the matching procedure, the benchmark should indeed minimize the incidence of Type II errors. We shall refer to this final set of Israeli patents as the "Benchmark Israeli Inventors Set," or ***BIIS*** for short.

VI.1 Using the BIIS to fine-tune the computerized matching procedure (CMP)

As already mentioned, contrasting the results of the CMP to the BIIS was one of the key methods used to try to improve the matching algorithm. The difficulty lied in the fact that there is no clear way of doing the comparison, let alone of quantifying it. In other words, any specific version of the CMP would render a list of unique Israeli inventors (and their corresponding patents), which obviously would not be identical to the BIIS – how could we then assess the "goodness of fit" between the two (if the latter is regarded as "data")? Spotty comparisons of differences between them are surely informative but can go only so far, and furthermore they cannot be too helpful if one considers multidimensional small changes in the matching parameters. We thus developed three alternative "goodness of fit indices", ***GOFIs***, and used them to fine tune the CMP *vis a vis* the BIIS: we adopted changes in the matching parameters that resulted in an improvement in these indices, worsening would lead to rejection of the changes, and mixed results would prompt us for further checks and close up examinations of the differences.

As a first stage, we "match" each unique inventor arrived at by the CMP (refer to it as ***"C"***) to its counterpart in the BIIS (call it ***"B"***). Accordingly, let $\boldsymbol{C_{ij}}$ be the set of all patents of inventor $\boldsymbol{j}$ named on patent (record) $\boldsymbol{i}$, as identified by the CMP, and $\boldsymbol{B_{ij}}$ the corresponding set found in BIIS. The indices are then defined as follows:

$$(1) \qquad GOFI_1 \equiv Mean\left[\frac{\left|B_{ij} \cap C_{ij}\right|}{\left|B_{ij} \cup C_{ij}\right|}\right], \quad i = 1,...,N_{IL}$$

where $\left|B_{ij} \cap C_{ij}\right|$ is the number of patents assigned to inventor $\boldsymbol{j}$ named in patent $\boldsymbol{i}$ both by the CMP and by BIIS, $\left|B_{ij} \cup C_{ij}\right|$ is the number of patents assigned to that inventor by the union of the two, and $\boldsymbol{N_{IL}}$ is the total number of patents/records associated with Israeli inventors. The idea is simply that we compute for each record of each inventor the share of the intersection of both sets out of the union of the sets: the max value is 1, which will be achieved only when both sets are exactly the same, and decreases as the two are less similar.

$$(2)a \quad GOFI_2 \equiv Mean\left[\frac{\left|B_{ij} \cap C_{ij}\right|}{\left|B_{ij}\right|}\right], \quad (2)b \;\; GOFI_2 \equiv Mean\left[\frac{\left|B_{ij} \cap C_{ij}\right|}{\left|C_{ij}\right|}\right]$$

The basic intuition is similar to that of $GOFI_1$, except that this index uses the number of patents assigned to the inventor by ***either*** method as the denominator, and not their union. In this case the comparison between (2)*a* and (2)*b* can be quite informative, in terms of which procedure is over or under matching relative to the other, and by how much. Thus for example if the CMP is under-matching then (2)*b* will be close to 1 and larger than (2)*a*.

We also developed a similarly constructed index to count the number of records handled differently by the two methods (note that there is no double counting of records matched differently). First define,

$$GOFI_3(i,j) = \begin{cases} 1 & [|C_{ij}| - |B_{ij} \cap C_{ij}|] \neq 0 \;\; or \;\; [|B_{ij}| - |B_{ij} \cap C_{ij}|] \neq 0 \\ 0 & otherwise \end{cases}$$

and,

$$(3) \qquad GOFI_3 \equiv \sum_{i,j} GOFI_3(i,j)$$

Thus $GOFI_3$ is the size of the file that contains all records that ended up matched differently by the two methods.

These indices allow us to diagnose the extent to which the CMP comes close to replicating the BIIS, which we regard as the "true" matching. In practice we proceeded as follows: first, we constructed the BIIS in parallel to developing the first-cut CMP; second, we tested, improved and refined the CMP in a variety of ways; lastly, we compared the (already much improved) CMP to the BIIS using the GOFI indices, and further fine-tuned the CMP. Table VI.1 shows the last round of the latter stage: as we can see, the two methods are quite "close" according to $GOFI_1$, but the difference in the values of $GOFI_2$ reveals that the CMP it still significantly under-matching. Indeed, the number of unique inventors identified by the CMP is 6,900, versus the 6,023 inventors singled out by the BIIS. This is a significant difference (of about 15%), and hence the question now is what accounts for such disparities.

Table VI.1
Comparing the CMP to the BIIS

	CMP	**BIIS**
Number of patents	9,155	
Number of records	15,306[29]	
Number of original names	6,314	
Number of Soundex-coded names (i.e. number of *p-sets)*	5,858	
Final number of unique inventors	6,900	6,023
Average number of patents per inventor	2.21	2.54
GOFI₁	0.88	
GOFI₂	0.99	0.89

[29] Ten "duplicate" records (i.e. records having the same name and same patent number) were deleted in the cleaning procedure.

Using $GOFI_3$ we identified the 5,081 records that were handled differently by the two methods, and proceeded to check the differences manually. The good news is that the incidence of Type II error induced by the CMP is indeed very low: there were only 73 inventors that the CMP over-matched (i.e. they corresponded to 196 inventors as identified by the BIIS). Furthermore, in most cases these were in fact not errors at all, but rather the CMP was right and thus the BIIS was wrong. Given that the emphasis in developing the CMP was in avoiding Type II error, it seems that goal was accomplished. The bad news is the high incidence of Type I error: the CMP under-matched in about 15% of cases, that is, it erroneously split 780 inventors into 1,781. The main reasons were:

1. ***Little in common*** *(or move without a trace)*: These are cases whereby two records turn out to refer to the same inventor, even though there is little or nothing in common between them other than the name. Formally, that means that the criteria used for matching failed to detect any similarity or linkage between the records. In these cases the matching of records by the BIIS was obviously done according to additional information not found in the patents themselves, and hence this is pretty much the upper bound of the matching ability of the CMP (or any such automated method that relies only on patent data).

2. ***Spelling mistakes in the names***: Soundex-coded names cannot overcome all possible spelling mistakes, and hence we may not match with the CMP two records that belong to the same inventor simply because they were not in the same *p-set* to begin with. This is a Type I error that could in principle be reduced if the coding improves.

3. ***Remaining errors in the spelling of cities and street addresses:*** As explained in Appendix 2, the quality of the match depends to a significant extent on the quality of the data fields used by the matching criteria. If of two records in a given *p-set*

one names "Jaffa" as the city of the inventor and the other "Yaffa", we probably will not match them even though we should.

Whereas the frequency of cases corresponding to cause 1 should be seen as an irreducible rate of Type I error, that is not so for causes 2 and 3: further cleaning of the data, and further fine-tuning of the Soundex method may significantly reduce these sources of Type I error as well. Close examination of the distribution of actual causes of Type I error revealed that about ½ of them correspond to cause 1, 1/3 of cases to cause 2 and the remainder of about 1/6 to cause 3. Thus, even if we were able to avoid Soundex-based and other spelling mistakes altogether, the CMP is still expected to result in ***7-8%*** of Type I errors, which thus constitutes a ***lower bound for Type I errors***. However, these results should be treated with caution, if only because Israeli inventors cannot be regarded as a random sample (in view of some of the peculiarities of Israeli names). For example, we know that for East-Asian inventors that situation is reversed: there is a high incidence of Type II errors relative to Type I. Clearly, further research experience is needed with these data in order to gain a better sense of its merits and limitations.

VII. The Average Matching Scores: Diagnostics and Fine-tuning

The end result of deploying the CMP of Stage 2 is a list of unique inventors, each with a number of records associated with her. As discussed in Section VI, one way of assessing the extent to which the CMP does a good job is by comparing it to the BIIS, but as said that represents just a small sample, and not necessarily a representative one. In order to assess the performance of the method in a comprehensive way and hence be able to further improve it, we rely on the average matching score (AMS) for inventor i, defined as:[30]

$$(4) \qquad AMS_i \equiv \frac{\sum_{j=1}^{m_i} pairwise\ \ score_j}{m_i}, \qquad m_i = \frac{N_i(N_i - 1)}{2}$$

[30] Note that the AMS is defined for inventors having more than one record, otherwise it is set to "missing."

where N_i is the number of records associated with inventor i (*after* applying the CMP), and m_i the number of all possible pair-wise comparisons of the records of that inventor. To exemplify, suppose that Stage 1 renders a *p-set* with 4 records (i.e. 4 patents that share the same Soundex-coded name), and that Stage 2 gives the following scores:

- Score (A, B): 310
- Score (B, C): 150
- Score (A, C): 80
- Score (D, *j*) < 100, j =A, B, C

Given the thresholds showed in Table V.1, records {A, B, C} would be grouped together as belonging to the same inventor (call him "John Fields"), and D becomes a separate inventor. Note that A and B are very similar, B and C less so, whereas A and C have little in common and in the end get matched just because of transitivity. We can now ask how "similar" to each other are on average the records {A, B, C} in so much as referring to the same inventor, or equivalently, how sure are we that John Fields as named in them is in fact one and the same person? Computing (4) for John Fields gives an average score of 150, which is quite high relative to the thresholds given in Table V.1. By contrast, suppose that the scores had been,

- Score (A, B): 120
- Score (B, C): 120
- Score (A, C): 0

In this case the AMS_i would be a mere 80, meaning that although we did group the three records together, they don't seem to have much in common (again, notice the role played by transitivity). In other words, the average probability of Type II error is thought to be inversely proportional to the corresponding AMS. Beyond serving as a diagnostic tool, the final AMS_i's could be used as indicators of the reliability of the respective observations in any econometric analysis, e.g. they could serve as weights in regression analysis. Having computed the AMS_i for each inventor i, we can then compute the overall matching score as,

$$(5) \qquad AMS = \frac{\sum_i N_i AMS_i}{N}, \qquad N = \sum_i N_i$$

i.e., AMS is the weighted average of the AMS_i, using the number of patents per inventor as weights, which allows us to assess the performance of the entire matching procedure.[31]

To recap: phase 1 of the development of the CMP consisted of a lengthy trial-and-error process by which a rudimentary algorithm was drafted (i.e. the "Mark I" CMP), in parallel to the creation of the benchmark Israeli inventors set (BIIS).[32] The second phase entailed a more systematic comparison of the Mark I CMP to the BIIS, with the consequent learning by doing and ensuing improvements to the CMP, which led to the "Mark II". At the end of phase 2 we had then a preliminary set of parameters (i.e. cutoff values, scores and thresholds) which allowed us to run the Mark II CMP, and obtain a (still tentative) list of unique inventors with their corresponding AMS's.[33]

In phase 3 we undertook a series of iterative "partial-derivative" changes, that is, in each step we changed just one parameter, run again the CMP (on a large sample of records) and analyzed the results of the change by comparing the new results to those obtained at the end of phase 2. Suppose for example that a change leads to significantly more matching decisions and correspondingly fewer distinct inventors (hence reducing the probability of Type I error), accompanied only by a minor reduction in the overall AMS (i.e. a minor increase in Type II error).[34] In such case we would be inclined to adopt the change, recalibrate the system accordingly, and restart the process by performing a marginal change in another parameter. In cases where the results were inconclusive, we examined manually a sample of records that were affected by the changed parameters so as to determine whether the changes were warranted or not.

---

[31] Given that the AMS is calculated on the basis of the scores used by the CMP, assessing the performance of the CMP and its sensitivity to different score values may be misleading: setting higher scores may lead to a higher AMS without real gains in reliability. Thus, the mean AMS should be judged against the maximal threshold value, and in any case it should not constitute the sole diagnostic tool.

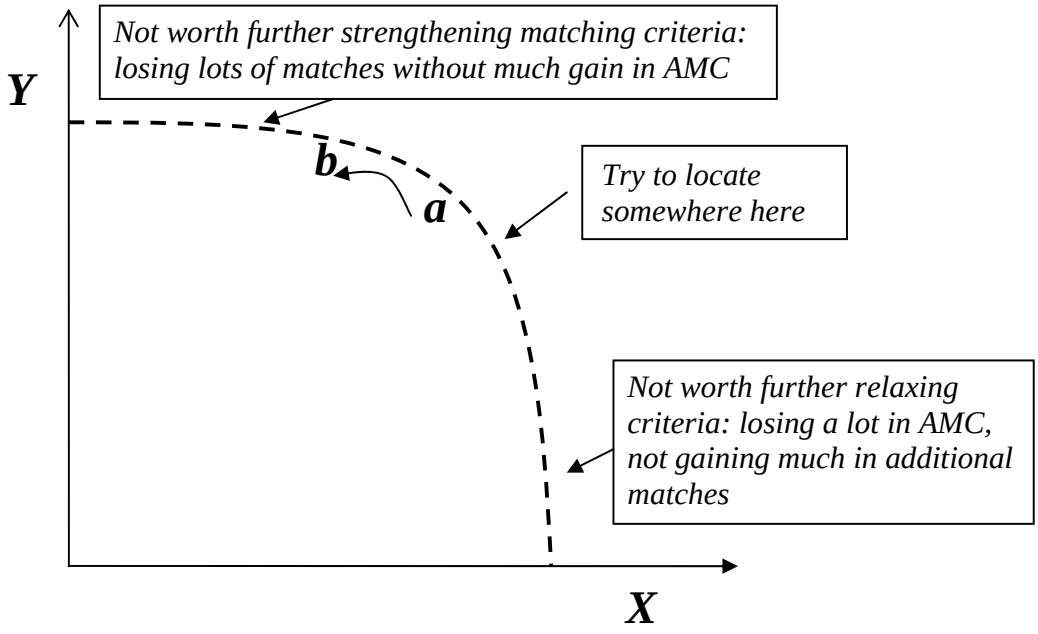
[32] By "phases" we mean here the actual steps taken in the process of developing the matching procedure; these are not to be confused with "stages" which reflect the logical segmentation of the process *ex post*.

[33] It is worth noting that the introduction of the AMS was a key "innovation" in developing the matching procedure: we essentially had stalled after what we refer to as "phase 2", and could not find ways of fine tuning the procedure in a systematic way, for lack of means to pierce the "black box" of millions of inventors with their records. The AMS offered precisely that ability (with the caveat of footnote 31), and hence allowed for "phase 3", which proved very effective.

[34] To avoid the potential fallacy alluded to in footnote 31, the AMS was calculated using the scores and cutoff values used *before* the partial-derivative change.

The basic trade-off involved is depicted in Figure 1, where ***Y*** stands for [1 – Prob(Type II error)], which is a (positive) function of the AMS, and ***X*** stands for [1 – Prob(Type I error)], which is inversely related to the number of unique inventors.[35] Suppose that we decrease the scores shown in Table V.3 (e.g. same Middle Name would get a score of 80 rather than 100), so that we end up with a smaller number of matches per inventor. Such a change would increase the AMS (given that it is now more difficult to get a match, if one does occur then the two records ought to be more "similar" to each other), and it would increase of course the number of unique inventors. In other words, the probability of Type II error decreases, but on the other hand the probability of Type I error increases. If the initial position along the transformation curve in Figure 1 was a point such as ***a*** and the change brings us to a point such as ***b***, the change was probably not worth its while: the loss in terms of Type I error is too large relative to the gain in the probability of Type II error.

Figure 1
Trade-Offs in Fine Tuning the CMP



[35] Note that inventors with just one record do not get an AMS value, and hence the overall mean AMS tells only part of the story: it should be always coupled with the number of unique inventors, as a diagnostic tool.

This is then the type of process that we followed, experimenting with dozens of such incremental changes, and thus mapping out the transformation curve schematically depicted in Figure 1. The good news is that the curve turned out to have indeed the shape shown, that is, beyond a certain mid area where the trade-offs posed a real dilemma, we quickly discovered that further movements in either directions were clearly unwarranted. On the other hand we can by no means claim that this process is sufficiently rigorous or comprehensive to render "optimal" parameters – there is surely room for further experimentation and improvements.

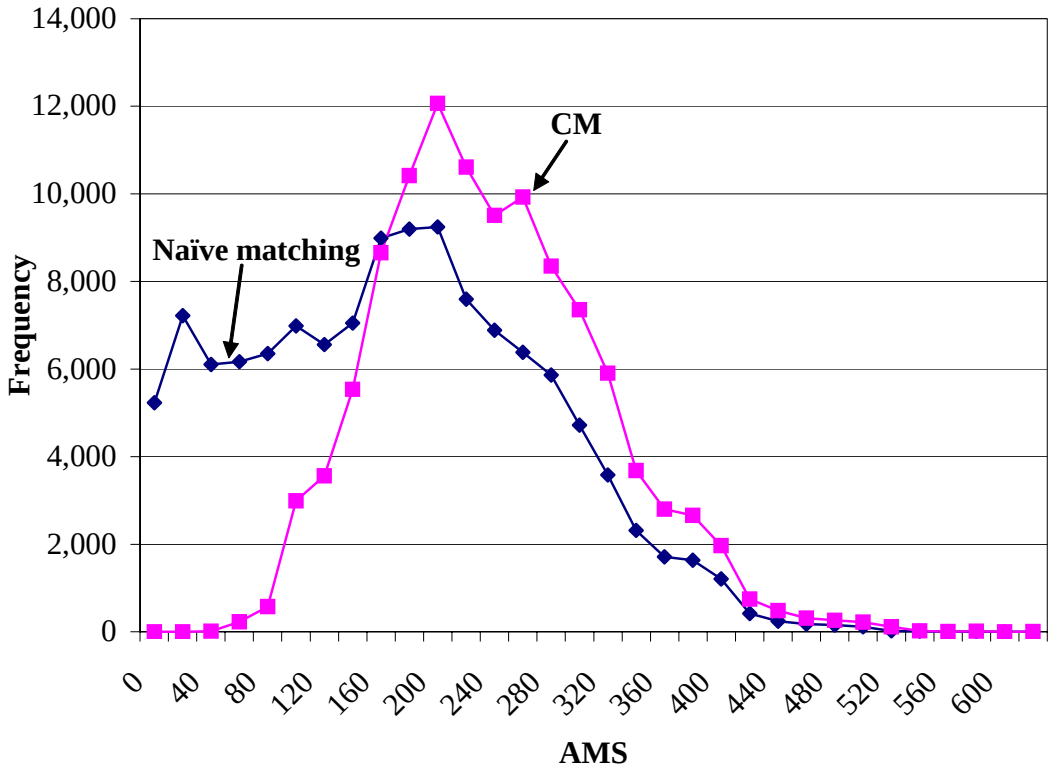
After converging at the end of phase 3 to a final set of parameters and producing the corresponding final list of unique inventors, we examined again the AMSs as a way of gaining further insight into the matching procedure. In particular, Figure 2 shows that the distribution of AMSs (over the set of inventors with more than one patent) is slightly skewed, and has a mean of 235. This implies that the average pair of records within a matched group satisfies either two "strong" criteria (worth a score of 120 each), or three weaker criteria, which is certainly a lot given that the CMP demands significantly less. Another encouraging implication of this result is that transitivity apparently does not play a significant role in the process, since if it did the AMS would likely be significantly lower.

Yet another angle at the performance of our matching procedure as reflected in the AMS is to pose the following question: suppose that we would have obviated altogether the development of a CMP, and had relied instead just on naïve matching, that is, we would have just grouped together records exhibiting exactly the same first and last inventor names. We could then compute for the ensuing list of unique inventors their AMS, and compare its distribution to the one that we obtained with our CMP. As Figure 2 shows, the naïve method shifts the distribution to the left, having a mean AMS of 171 versus 235 for ours.[36] Moreover, a large number of inventors in the naïve case get an AMS of zero, meaning that pairs of records presumed to belong to the same inventor do

[36] For simplicity of computation, the comparison was done only for a sample of inventors, those with last names that begin with the letter 'A'.

not have anything in common (according to the criteria used here). Thus, there is reason to believe that our CMP does improve significantly the matching, and that there are no easy shortcuts.

Figure 2
Comparing the CMP to "naïve" matching*



* In order to reduce the computational burden, the naïve matching was done just for a sample of inventors: those with family names starting with the letter "A".

VIII. The final list of unique inventors: summary figures

As mentioned in Section II, we started from the NBER Patent and Citations Data File that contains 2,139,313 patents covering the period 1975-1999, and the corresponding Inventors file. Defining each inventor in each patent as a "record", these data generated 4,298,457 records, which implies that the average number of inventors per patent is 2.01. As a result of deploying our matching procedure we obtained a final list of 1,632,532 unique inventors, assigned to them an ID, and merged this list back to the 4.3

million records. Thus, we now know who invented what, with whom, where, and so forth. On that basis we can trace the histories of those inventors, follow their movements across assignees, geographic areas and technological fields, explore the determinants of their innovative outputs as a function of their previous experience, etc. This is an extraordinarily rich throve of data, that opens a wide range of research possibilities.

In order to appreciate what the CMP developed here does, consider the following names that appear very frequently in our data, and the number of different inventors in which they were split at the end of the procedure:

- Robert Smith (749 records) – 271 IDs
- David Smith (643 records) – 227 IDs
- Robert Miller (588 records) – 176 IDs

Thus for example, there are 749 patents in which the (exact) name Robert Smith appears; applying to them the CMP rendered a list of 271 ***different*** inventors having that name. Had we not done that and relied instead on "naïve matching", we would have ended up with one seemingly extremely prolific inventor, whereas in reality Robert Smith "is" 271 different individuals.

It is reassuring to note that the recent USA Today's list of top 10 living US patent holders,[37] is supported by our results, even though our data are as of 1999 and hence comprise only a fraction of the total as of 2005. As Table VIII.1 shows, there is a significant overlap between the two rankings, certainly for the "older" inventors (i.e. inventors whose first patent was granted in the 1970s).

[37] http://www.usatoday.com/tech/columnist/kevinmaney/2005-12-13-patent_x.htm

Table VIII.1
USA Today Top 10 Living US Patent Holders
For US Today # of patents as of 12.13.2005; our data up to 12. 1999

Inventor name	US Today	Our data	Our ranking	Grant year of 1st patent[38]
1. Shunpei Yamazaki	1,432	605	3	1979
2. Donald Weder	1,322	466	4	1978
3. Kia Silverbrook	801	58	1652	1994
4. George Spector	723	715	1	1976
5. Gurtej Sandhu	576	172	89	1991
6. Warren Farnworth	547	128	216	1990
7. Salman Akram	527	93	450	1995
8. Mark Gardner	512	233	37	1994
9. Heinze Focke	508	388	8	1976
10. Joseph Straeter	477	133	193	1991

In order to gain some perspective of what the number of unique inventors that we arrived at means, consider the following bounds: the starting point was that there could be in principle as many unique inventors as there are records, i.e. 4,298,457; this is so since any two names, even if identical, could refer in fact to different individuals. At the other end, if we were to be extremely lenient and match any two inventors just on the basis of having the same Soundex-coded name, we would have 630,887 distinct inventors. Being slightly more demanding and matching on the basis of having just identical first and last names would render 1,205,403 inventors, or 1,405,318 inventors if using in addition the middle name initial. As can be seen in Table VIII.2, deploying the CMP makes indeed a big difference *vis a vis* the "naïve" alternatives in both directions in terms of raw numbers. Furthermore, it is not just that the CMP renders 1.6 million inventors versus say the 1.2 million obtained by matching according to identical names:

[38] The year of first patent was taken from the USPTO, and in it the data starts from 1976, hence when it says 1976 it may have been an earlier year.

the lists are different since the latter are not necessarily a sub-set of the former (due to the use of Soundex-coded names).

Table VIII.2
Matching in Perspective

Matching method	**Number of unique inventors**
Each record a different inventor	4,298,457
Computerized matching procedure (CMP)	**1,632,532**
Identical last and first names and middle name initial	1,405,318
Identical last and first names	1,205,403
Same Soundex-coded names	630,887

Although there are surely many more inventors that do not ever patent, our list of 1.6 million inventors are quite likely responsible for the vast majority of innovations made over the last 3 decades of the 20th century, almost certainly for the important ones. As Table VIII.3 reveals, the distribution of number of patents per inventor is very skewed (as virtually everything else regarding patents), with an average of 2.6 patents per inventor. For purposes of research on mobility and careers of inventors, the interesting data are those related to the 0.7 million inventors that have at least 2 patents – the 1.0 million with just one patent (the "occasional" inventors) are certainly important but obviously cannot shed light on those research issues. Notice that there are about 70,000 with more than 10 patents: those are the most prolific inventors, and they will probably command a great deal of attention in coming research.

Table VIII.3 Distribution of Patents per Inventor		
Patents per inventor	**Number of Inventors**	**% of inventors**
1	983,859	60.27
2 - 5	497,780	30.49
6 - 9	80,835	4.95
10 - 50	67,565	4.14
50+	2,402	0.15
total	***1,632,441***	***100.00***

References

Alcacer, Juan and Gittelman, Michelle, “How do I know what you know? The role of inventors and examiners in the generation of patent citations.” Cross Disciplinary Strategy Seminar, Stern School of Business, NYU, Spring 2004.

Algawar, Ajay K., Cockburn, Iain M., McHale, John, “Gone But Not Forgotten: Labor Flows, Knowledge Spillovers and Enduring Social Capital.” National Bureau of Economic Research Working Paper No. 9950, September 2003.

Breschi, Stefano and Lissoni, Francesco, "Mobility and Social Network: Localized Knowledge Spillovers Revisited", CESPRI Working Paper No. 142, March 2003.

Crespi, Gustavo A., Geuna, Aldo and Nesta, Lionel J. J., "Labor Mobility of Academic Inventors. Career Decision and Knowledge Transfer", EUI Working Paper RSCAS No. 2006/06, June 2006.

Fleming, Lee and Marx, Matt, “Non-competes and inventor mobility: the Michigan Experiment.” Harvard Business School Working Paper, 2006.

Griliches, Zvi, (ed.) R&D, Patents, and Productivity, NBER Conference Proceedings. University of Chicago Press, 1984.

Griliches, Z., Hall, B.H. and A. Pakes, "The Value of Patents as Indicators of Inventive Activity," in P. Dasgupta and P. Stoneman, eds., Economic Policy and Technological Performance. Cambridge: Cambridge University Press, pp. 97-124, 1987.

Griliches, Zvi, “Patent Statistics as Economic Indicators,” Journal of Economic Literature 92: 630-653, 1990.

Hoisl, Karin, "Tracing Mobile Inventors – The Causality between Inventor Mobility and Inventor Productivity", Munich Business Research Working Paper Series No. 2006-9, May 2006.

Jaffe, A., Trajtenberg, M. and R. Henderson, “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations,” Quarterly Journal of Economics, pp. 577-598, August 1993.

Jaffe A. and Trajtenberg, M, Patents, Citations and Innovations: A Window to Knowledge Economy. Cambridge Mass: MIT Press, 2002.

Jones, Benjamin F., “The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?” National Bureau of Economic Research Working Paper No. 11360, May 2005.

Kim Jinyoung, Lee Sangjoon John, Marschke Gerald, “The Influence of University Research on Industrial Innovation.” National Bureau of Economic Research Working Paper No. 11447, June 2005.

Pakes, Ariel and Simpson, Margaret, “The Analysis of Patent Renewal Data.” Brookings Papers on Economic Activity, Microeconomic Annual, pp. 331- 401, 1991.

Rosenkopf, Lori and Almeida, Paul, “Overcoming Local Search Through Alliances and Mobility.” Management Science, Vol. 49, No. 6, pp. 751-766, June 2003.

Schankerman, M. and A. Pakes, “Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period,” Economic Journal, Vol. 96, No. 384, pp. 1052-1077, December 1986.

Scherer, F.M. “Inter-Industry Technology Flows and Productivity Growth,” Review of Economics and Statistics, 64, November 1982.

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

Singh, Jasjit, "Inventor Mobility and Social Networks as Drivers of Knowledge Diffusion", Harvard University Working Paper Series, October 2003.

Song, Jaeyong, Almeida, Paul and Wu, Geraldine, “Learning-by-Hiring: When is Mobility More Likely to Facilitate Inferfirm Knowledge Transfer?.” Management Science, Vol. 49, No. 4, pp. 351-365, April 2003.

Stolpe, Michael, “Mobility of Research Workers and Knowledge Diffusion as Evidence in Patent Data – The Case of Liquid Crystal Display Technology.” Kiel Working Paper No. 1038, April 2001.

Trajtenberg, M. "A Penny for Your Quotes: Patent Citations and the Value of Innovations," The Rand Journal of Economics, 21(1), 172-187, Spring 1990.

Trajtenberg, M. “The Names Game: Using Inventors Patent Data in Economic Research”. **http://www.tau.ac.il/~manuel/**, Power-point presentation, 2004.

Trajtenberg, M., Jaffe, A. and R. Henderson, "University versus Corporate Patents: A Window on the Basicness of Invention," Economics of Innovation and New Technology, 5 (1), pp. 19-50, 1997.

Zucker, Lynne G., and Darby, Michael R., "Movement of Star Scientists and Engineers and High-Tech Firm Entry", National Bureau of Economic Research Working Paper No. 12172, April 2006.

Appendix 1
Example of a Front Page of a Patent

United States Patent **4,203,158**
Frohman-Bentchkowsky, et. al. **May 13, 1980**

Electrically programmable and erasable MOS floating gate memory device employing tunneling and method of fabricating same

Abstract

An electrically programmable and electrically erasable MOS memory device suitable for high density integrated circuit memories is disclosed. Carriers are tunneled between a floating conductive gate and a doped region in the substrate to program and erase the device. A minimum area of thin oxide (70 A-200 A) is used to separate this doped region from the floating gate. In one embodiment, a second layer of polysilicon is used to protect the thin oxide region during certain processing steps.

Inventors: **Frohman-Bentchkowsky; Dov** (Haifa, IL); **Mar; Jerry** (Sunnyvale, CA); **Perlegos; George** (Cupertino, CA); **Johnson; William S.** (Palo Alto, CA).
Assignee: **Intel Corporation** (Santa Clara, CA).
Appl. No. **969,819**
Filed: **Dec. 15, 1978**

Related U.S. Application Data

Continuation-in-part of Ser No. 881,029, Feb. 24, 1978, abandoned.
Intl. Cl. : **G11C 11/40**
Current U.S. Cl.: **365/185.29**; 257/321; 326/37; 327/427;
Field of Search: **365**/185, 189; **307**/238; **357**/41, 45, 304

References Cited | [Referenced By]

U.S. Patent Documents

3,500,142	Mar., 1970	Kahng	**365**/185
4,051,464	Sept., 1977	Huang	**365**/185

Primary Examiner: Fears; Terrell W.

16 Claims, 14 Drawing Figures

Appendix 2
Cleaning and Standardizing the Raw Data

As described in Section IV, the matching procedure relies *inter alia* on pair-wise comparisons of the values of data fields such as the inventor's addresses, city, state and country. Clearly, a necessary condition for the comparisons to be meaningful is that these values are error-free, otherwise one cannot tell whether two strings differ because they refer to different entities or because some of them contain erroneous characters. Therefore, first of all we had to clean up and standardize the alphanumeric fields that proved to be particularly prone to error: street address, city and state. Still, the cleaning process that we implemented cannot be regarded as "complete", and was intended to solve primarily the most frequent and severe cases; there is certainly room for further improvement in this dimension that may prove cost-effective.

Street address

For street addresses we simply changed all characters to upper-case letters and removed most of the non-alphanumeric characters (i.e. parenthesis, apostrophes, commas, &, **:**, /, `, **;** , etc.). This is by no means a comprehensive standardization procedure: the street number may appear before or after the street name, the street name may or may not be followed by the abbreviation "St." or "Ave.", etc. As to the cost-effectiveness of further "cleaning": on the one hand street addresses appear only in 11% of the records, the number of variations in format is quite large, and the records will be anyway scored for the city criteria. On the other hand street address is one of the three strongest criterion for matching, and hence missing on it may be consequential.[39] More scrutiny is needed here to assess whether or not further work would be worthwhile.

City name

As apposed to street addresses, virtually all records contain the city of the inventor, and city co-location does play an important role in the matching procedure. Furthermore, we have refined its use as a matching criterion so as to take into account city size: location in larger cities (according to the number of patents) is given a lower score than small towns. Therefore, we invested significant efforts to clean up and standardize city names;[40] in addition to changing the names to upper-case letters and removing the non-alphabetic characters, we identified and standardized the following frequent occurrences:

- The name of the city appears as part of a string that may include the country code, state (or province), and ZIP, with or without parenthesis, in which case we just extracted the city name and deleted the rest. Here are some examples: "B-8791 Waregem-Bever", "CH-8103 Unterengstri", "Fabreville, Quebec", "64700, Monterrey, N", "Fano (Pesaro)", "Cortaillod (Ne)" or "Berthierville (Quebec)".

[39] If two records with the same Soundex coded name share only the same address, the full address criterion might be sufficient for a match, while the city name criterion will not be sufficient.

[40] Note that Soundex is not appropriate for city names: because the USPTO scans the data using optical-character recognition, the errors are likely to stem from glitches in the OCR software, and not from phonetic misrepresentation (see Fleming & Marx, 2006).

- City names that include variations of "Saint" were standardized to the "St" prefix (e.g. St Louis instead of Saint Louis).
- The names of major non-US cities were spelled according to their English version, e.g. Rome and not Roma, Milan and not Milano, Munich and not Muchen, etc.

The importance of cleaning cities names can be seen in the fact that the number of different cities shrank from 177,696 at the start to 133,282 after cleaning.

State and Country

Even though countries and US states appear in the original patent files as a 2-letter code and in principle do not require addition "cleaning," we discovered a non-negligible number of records for which state codes and (identically written) country codes were mistakenly interchanged (e.g. CA may stand for California or for Canada, PA for Pennsylvania or for Panama, etc.). A few initial cases that came to our attention alerted us to the potential problem, which is not easy to treat in a systematic way (e.g. one cannot possibly check every California patent to see whether it refers perhaps to Canada). We adopted instead a pragmatic (but limited) approach as follows:[41]

- We identified all patents designated as Canadian that contained cities that are known to be in California, and in particular, cities that include in their names the words "Rancho", "San" or "Palo"; those patents were reassigned to California.

- Any patent assigned to Panama as a country (PA code) that was ***not*** assigned to a city that includes "Panama" in its name (mostly Panama-City) was changed to Pennsylvania as a state and to the US as country.

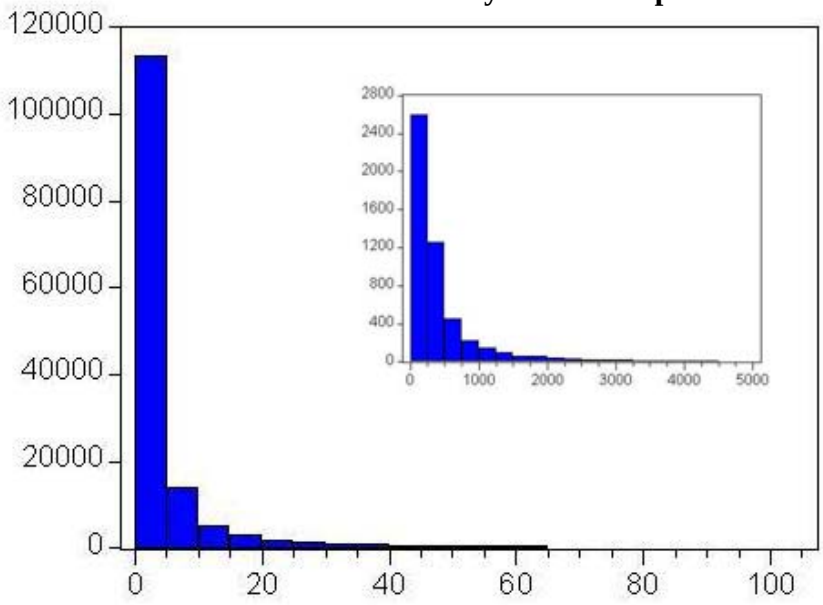
[41] There may remain further problems with other country-state codes (e.g. Israel – Illinois, both coded IL), but we did conduct extensive checks and did not find systematic errors.

Appendix 3
Size and Frequency-Dependent Categories: Names, Cities, Assignees and Patent Classes

As explained in Section IV, the scores of some of the matching categories are made to depend upon the "size" of those categories, since having two records say in the same city is more informative the smaller the city is, and likewise for assignee and for patent class. Likewise, the less frequent a Soundex-coded name is, the more informative is the fact that two records bear the same Soundex code.

In order to determine what should be regarded as "small" or "large", we examined the distribution of cities according to number of patents, and looked for appropriate cutoff values, such as first moments (for ease of exposition we shall refer to cities as the paradigmatic case, but the discussion applies equally to the other categories as well). As shown in Figure 3 though, the distributions turned out for the most part to be extremely skewed, and hence neither the mean nor the median seemed able to offer reasonable cutoff values (they were far too low).

Figure 3
Distribution of Cities by number of patents



Thus, we looked instead at the distribution of the number of *records* (patents) by *city size* rather than at the number of cities by city size, and examined the median values. To make it clear, define x**:** number of patents in a city (i.e. its "size"), $f(x)$ number of cities of size x, and $y = x\,f(x)$**:** number of patents in cities of size x; the parameter of interest is the median value of the distribution of y. We then eyeballed the cities around this median value, to see whether those above it could indeed be regarded as sufficiently "large", and conversely for those below. In the case of cities and patent classes, these

median values turned out to be reasonable cutoff values, and thus we adopted them. In the case of assignees a lengthy experimentation and manual examination process led us to reduce the cutoff value significantly, and similarly for Soundex-coded names. We now provide further detail on each of the matching criteria and the corresponding distributions.

A3.1 Cities

Table A3.1.1
Distribution of Cities, by "Size"

Number of records (city "size")	Number of cities	Total # of records by city size
1	64,394	64,394
2	18,306	36,612
3	89,65	26,895
4	5,541	22,164
5	3,946	19,730
6	3,061	18,366
7	2,357	16,499
8	1,953	15,624
9	1,526	13,734
10	1,365	13,650
11-20	7,264	106,641
21-30	3,246	81,157
31-50	3,110	121,887
51-100	3,135	224,013
101-300	3,506	704,582
301-1,000	1,235	865,240
1,001-1,382	7,264	106,641
Cutoff: 1382[42]		
1,383-5000	3,66	852,811
5,001-10,000	58	393,970
10,001-50,000	29	490,003
50,000+	2	210,487

Table A3.1.2 presents the ten "largest" cities in our data, that is, the cities with the largest number of patents originating in them. Note that the top five are Japanese: this is in part an artifact of how city limits are drawn in different countries (e.g. they are much more encompassing in Japan than in the US), but it also reflects a real phenomenon,

[42] This corresponds to the median of the distribution computed in a previous round; the median has slightly changed since, but the change is immaterial and hence we left the cutoff as is.

namely, that innovative activity in Japan is much more geographically concentrated than in the US. One of the consequences is that city location is less informative for matching of Japanese inventors than for American-based inventors, a fact that only compounds the problems associated with East-Asian names.

Note also that Austin (Texas) has more patents than New York City (which does not appear in the table, since it occupies the 20nd place with only 12,840 patents), but again this reflects to a large extent different municipal designs. In fact, there are over 6,178 patents assigned to the Bronx, Queens, Brooklyn and Staten Island, which are essentially part of New York City. Similarly for Los Angeles, which stands at the 39th place with 9,111 patents.

Table A3.1.2
10 "Largest" Cities
(i.e. cities with the largest number of records)

City	Country	# of records ("size")
Tokyo	Japan	135,910
Yokohama	Japan	74,577
Kanagawa	Japan	47,695
Kawasaki	Japan	40,615
Osaka	Japan	33,360
Houston (TX)	USA	26,241
San Jose (CA)	USA	22,573
Rochester (NY)	USA	18,452
Austin (TX)	USA	17,910
Saitama	Japan	16,768

A3.2 Assignees

Table A3.2.1
Distribution of Assignees, by "Size"

Number of records (assignee "size")	Number of assignees	Total # of records by assignee size
1	49,987	49,987
2	31,231	62,462
3	15,256	45,768
4	10,846	43,384
5	5,986	29,930

6	5,291	31,746
7	3,466	24,262
8	3,035	24,280
9	2,266	20,394
10	2,014	20,140
11-20	6,281	79,963
21-30	3,387	60,245
31-50	6,922	216,917
51-100	2,565	181,613
101-300	1,988	333,970
301-1,000	897	476,684
1,001-1,540	156	193,396
Median: 1540		
1,540-5000	219	586,298
5,001-10,000	53	360,722
10,001-50,000	39	760,762
50,000+	4	695,534

Cutoff of 500

The 10 largest assignees are shown in Table A3.2.2: not surprisingly the list comprises the usual suspects, i.e. major corporate firms that do also the most industrial R&D. Note that half of them are Japanese, pointing again to the high concentration of innovation in Japan, both geographically and firm-wise.

Table A3.2.2
10 Largest Assignees

Assignee	**Base country**	**# of patents ("size")**
HITACHI, LTD	Japan	70,921
IBM, LTD	USA	63,311
CANON KABUSHIKI KAISHA	Japan	52,994
GENERAL ELECTRIC COMPANY	USA	38,297
BAYER AKTIENGESELLSCHAFT	Germany	37,200
TOSHIBA CORPORATION	Japan	36,290
MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.	Japan	32,316
MITSUBISHI DENKI KABUSHIKI KAISHA	Japan	30,604
BASF AKTIENGESELLSCHAFT	Germany	27,806
EASTMAN KODAK COMPANY	USA	27,720

A3.3 Patent Classes

Table A3.3.1 Distribution of Patent Classes, by "size"		
Number of records (Pat class "size")	**Number of patent classes**	**Total # of records by pat class size**
2-10	6	29
11-100	7	271
101-200	7	1,165
201-500	17	5,913
501-1000	20	14,300
1001-2000	36	50,015
2001-5000	88	303,041
5001-10000	94	677,744
10001-18861	81	1,117,761
Median: 18,861 – cutoff		
18862-25000	22	474,230
25001-50000	33	1,138,553
50000+	6	515,435

Table A3.3.2 Ten Largest Patent Classes		
Patent class code	**Patent class**	**# of records**
514*	Drug, bio-affecting and body treating compositions	163,051
428	Stock material or miscellaneous articles	90,736
435	Chemistry: molecular biology and microbiology	76,919
430	Radiation imagery chemistry: process, composition, or product thereof	68,628
424	Drug, bio-affecting and body treating compositions	61,405
73	Measuring and testing	54,696
123	Internal-combustion engines	49,513
257	Active solid-state devices (e.g., transistors, solid-state diodes)	46,379
438	Semiconductor device manufacturing: process	45,832
250	Radiant energy	44,521
** Part of class 424*		

A3.4 Soundex-coded names

Table A3.4.1 **Distribution of Soundex-Coded** *("S-coded")* **Names, by "Size"**		
# of records (S-coded name "size")	**# of S-coded names**	**Total # of records by S-coded names size**
1	257,904	257,904
2	100,598	201,196
3	56,748	170,244
4	37,454	149,816
5	26,947	134,735
6	19,991	119,946
7	15,571	108,997
8	12,350	98,800
9	10,364	93,276
10	8,689	86,890

11-13	19,323	229,707
14-16	12,829	191,330
Cutoff: 16		
17-20	11,635	213,922
21-30	15,769	391,972
31-50	12,444	481,532
51-100	8,154	561,357
101-500	3,958	698,545
501-1000	151	97,248
1000+	8	11,040

Median: 23

Appendix 4
East-Asian Names

As mentioned in Section IV, using Soundex for coding names that are not English or German-based might not be appropriate, and in particular Soundex may unduly increase the probability of Type II errors when deployed on East-Asian names. By East-Asians we mean inventors having as country of residence China, Hong Kong, Japan, Korea, Singapore or Taiwan. Of course, there are many American inventors of East-Asian origin that may display oriental names, and ideally we would like to treat differently these as well, since the problem is linguistic, not geographic. However, identifying names linguistically (or ethnically) is exceedingly difficult, and hence we have resorted to the expedient device of singling out oriental names as they occur with inventors located in East-Asia.

The key issue with East-Asian names is that they typically include many vowels relative to consonants as well as frequent appearances of the letter H, whereas the Soundex code ignores both and hence a significant part of the name information gets lost. In addition, there is a high incidence of short names. Thus in many cases the Soundex code for East-Asian names consists just of the initial followed by very few non-zero digits: as shown in Table A4.1, a full 75% of Korean family names, 65% of Taiwanese and 58% of Chinese either have just the initial or a single digit following it, and hence for these names the Soundex code is largely non-informative.[43] Japanese names also suffer from the same syndrome but to a lesser extent. As Table A4.2 reveals, the differences are much less pronounced for first names, and indeed, as indicated below, we will rely heavily on them. Notice that on average European and American Soundex-coded names are over 2.5 times longer than Korean, Taiwanese and Chinese names, and about 50% longer than Japanese names. Thus using the Soundex-code for East-Asian names induced a high incidence of Type II errors, that is, we tended to match together records that in fact belonged to different inventors.

Further evidence of the same phenomenon is given in Table A4.3, showing the 10 most frequent Soundex-code names (largest *p-sets)* in the entire data: ***all*** of them happen to be East-Asian. Since several names are typically coded into the same Soundex code, Table A4.3 shows also the three most frequent names for each Soundex code, and the number of records associated with them (in parenthesis). Note once again that since vowels as well as the letters H, W and Y are ignored, the Soundex code turns out to be very short for names that are rather long (e.g., TAKAHASHI HIROSHI is coded into T220000 H620000, where the zeros are non-informative), which eventually might cause Type II errors (e.g., it can be matched with last names such as TSUYUKI, TAKAGI TAJIKA etc.).

[43] It is worth noting though that there were very few inventors located in China in that period (just 2,416), and hence the figures for China are not very informative.

Table A4.1
Percentage Distribution of Inventors, by # of Digits of *Last* Name and Country
(sorted by % of 0-1 digits)

Country	Number of digits			
	0 – 1	**2**	**3 or more**	**Mean**
Korea	**75.6**	**23.8**	**0.6**	**0.89**
Taiwan	**65.5**	**30.5**	**4.0**	**1.05**
China	**59.3**	**33.2**	**7.5**	**1.10**
Japan	**29.2**	**47.3**	**23.5**	**1.96**
US	17.6	32.5	49.9	2.58
Canada	17.3	33.0	49.8	2.58
Britain	16.6	34.4	49.0	2.57
Israel	14.4	38.2	47.4	2.66
Italy	13.9	37.4	48.7	2.58
Germany	12.9	30.5	56.6	2.80
France	12.9	37.2	49.9	2.61
Netherlands	9.4	20.4	70.2	3.29

Table A4.2
Percentage Distribution of Inventors, by # of Digits of *First* Name and Country
(sorted by % of 0-1 digits)

Country	# of digits			
	0 – 1	**2**	**3 or more**	**Mean**
China	**37.8**	**33.1**	**29.1**	**1.84**
Israel	32.4	45.4	22.2	1.94
Korea	**31.9**	**40.4**	**27.8**	**1.90**
Taiwan	**29.6**	**28.6**	**41.8**	**2.27**
Japan	**29.1**	**54.0**	**16.9**	**1.88**
Canada	22.1	40.8	37.1	2.23
US	21.5	42.1	36.4	2.19
Italy	20.0	39.8	40.2	2.38
France	19.4	38.0	42.5	2.42
Britain	19.3	47.0	33.6	2.23
Netherlands	16.9	31.4	51.7	2.71
Germany	5.8	33.0	61.3	2.90

Table A4.3
The 10 Soundex-Coded Names with the Largest Number of Records

Soundex code	3 most frequent names belonging to this Soundex code	# of Records
S300000 K200000	SATO KOICHI (137) SATO KOZO (130) SATO KAZUO (123)	1753
S300000 T220000	SATO TAKASHI (147) SAITO TAKASHI (131) SATO TAKESHI (92)	1506
T220000 H620000	TAKAHASHI HIROSHI (247) TAKEUCHI HIROSHI (143) TAKAHASHI HIROYUKI (119)	1470
T220000 K200000	TAKAHASHI KOJI (185) TAKAHASHI KOICHI (100) TAKAHASHI KAZUO (82)	1453
S220000 T220000	SUZUKI TAKASHI (316) SUZUKI TAKESHI (160) SASAKI TAKASHI (145)	1424
S300000 H620000	SATO HIROSHI (277) SAITO HIROSHI (119) SAITO HIROYUKI (89)	1208
S300000 Y200000	SATO YUICHI (74) SUDA YASUO (68) SATO YOSHIO (67)	1134
T220000 T220000	TAKAHASHI TAKESHI (86) TAKEUCHI TAKASHI (76) TAKAHASHI TAKASHI (58)	1092
I300000 H620000	ITO HIROSHI (229) ITOH HIROSHI (160) IWATA HIROSHI (137)	985
S220000 K200000	SUZUKI KOJI (117) SUZUKI KAZUO (102) SUZUKI KOICHI (91)	975

These findings led us to alter the matching procedure as follows: we flagged East-Asian names, and imposed as an additional necessary condition for a match to occur between any two such records that the original first name be exactly the same. There are probably better ways to handle East-Asian names, and moreover, whatever the procedure adopted, it surely should be applied to all inventors bearing such names and not just to

those having an address in East-Asian countries. As said, the Soundex system is by and large not appropriate for those names, and we hope that better coding systems would be developed in the future.

Lastly, it is worth reporting the procedure by which we discovered and assessed the nature and prevalence of the problem with East-Asian names, since in so doing some interesting facts arise. Originally we had deployed the Soundex code only on the last name, not on the first name. However, in the course of calibrating the CMP to the BIIS, we found that in a non-negligible number of cases there were slight differences in the spelling of the first names, which precluded the matching of records that should be matched. Thus we decided to extend the Soundex code to include first names as well. However, we then discovered that this change may be inducing Type II errors in East-Asian names, and proceeded to investigate this further. We computed the frequency of records by letters as well as the incidence of East-Asian names in each letter, and found that "R" has the lowest percentage of oriental names, among the top half of the letters (see Table A4.4).[44] We then applied the CMP to the records in R with and without extending the Soundex to the first name, and compared the results: the overwhelming majority of matching errors induced by Soundex-coded first names occurred indeed in oriental names. We then repeated the exercise with the letter L (6% of East-Asian names), and with the letter I, which has the highest percentage of oriental names (a staggering 84%), and found the same result: most errors occurred with oriental names.

Table A4.4
Distribution of Records by Letter:
Total and East-Asian Inventors
(sorted by # of records)

Letter	Number of records		% East-Asian
	Total	East-Asian	
S	487,399	133,568	27.4
M	370,205	108,807	29.4
K	342,610	159,994	46.7
H	305,890	80,399	26.3
B	291,241	3,851	1.3
T	219,353	114,766	52.3
C	205,371	13,439	6.5
W	190,755	20,487	10.7
L	183,539	11,063	6.0
G	172,762	5,456	3.2
P	160,462	2,974	1.9
R	158,819	1,133	0.7
F	153,409	35,042	22.8

[44] The letter V has even fewer oriental names, but then it is a relatively infrequent letter altogether, and hence we decided to check R instead.

D	152,947	4,033	2.6
N	150,371	82,973	55.2
A	141,290	43,597	30.9
O	120,795	76,242	63.1
Y	105,034	82,913	78.9
I	98,459	82,847	84.1
J	76,213	3,995	5.2
E	71,879	9,221	12.8
V	66,827	91	0.1
U	35,512	24,300	68.4
Z	31,273	1,057	3.4
Q	4,943	60	1.2
X	1,099	151	13.7
Total	**4,298,457**	**1,102,459**	**25.6**

Appendix 5

Computing the Goodness of Fit Indices (GOFIs): some examples

Example 1: The CMP did not match any record (hence each received a different ID), whereas the BIIS matched the 3 of them. Note that $GOFI_2$=2 for the CMP, whereas it equals just 1/3 for BIIS, indicated that the CMP significantly under-matches.

Example 1 **Inventor Name: Almagor David**										
Patent	**CMP ID**	**BIIS ID**	$\lvert B_{ij} \rvert$	$\lvert C_{ij} \rvert$	$\lvert B_{ij} \cap C_{ij} \rvert$	$\lvert B_{ij} \cup C_{ij} \rvert$	***GOFI₁***	***GOFI₂***		***GOFI₃***
								BIIS	CMP	
4862427	1084474	329	3	1	1	3	1/3	1/3	1	1
5541967	1084475	329	3	1	1	3	1/3	1/3	1	1
5631957	1084476	329	3	1	1	3	1/3	1/3	1	1
Mean							1/3	1/3	1	3 *(total)*

Example 2: the two methods render exactly the same matching.

Example 2 **Inventor Name: Rotem Eran**										
Patent	**BIIS ID**	**CMP ID**	$\lvert B_{ij} \rvert$	$\lvert C_{ij} \rvert$	$\lvert B_{ij} \cap C_{ij} \rvert$	$\lvert B_{ij} \cup C_{ij} \rvert$	***GOFI₁***	***GOFI₂***		***GOFI₃***
								BIIS	CMP	
5684412	14261	18109639	4	4	4	4	1	1	1	0
5903490	14261	18109639	4	4	4	4	1	1	1	0
5861641	14261	18109639	4	4	4	4	1	1	1	0
5751165	14261	18109639	4	4	4	4	1	1	1	0
Mean							1	1	1	0

Example 3: The two methods differ just in one record (the last one).

Example 3 **Inventor Name: Amit Noah**										
Patent	**BIIS ID**	**CMP ID**	$\lvert B_{ij} \rvert$	$\lvert C_{ij} \rvert$	$\lvert B_{ij} \cap C_{ij} \rvert$	$\lvert B_{ij} \cup C_{ij} \rvert$	***GOFI₁***	***GOFI₂***		***GOFI₃***
								BIIS	CMP	
5120225	523	1097272	4	3	3	4	3/4	3/4	1	1
5186627	523	1097272	4	3	3	4	3/4	3/4	1	1
5494252	523	1097272	4	3	3	4	3/4	3/4	1	1
4923460	523	1097275	4	1	1	4	1/4	1/4	1	1
Mean							**0.625**	**0.625**	**1**	**4** *(total)*

Appendix 6
Examples of Matching with the CMP

Example 1:
Strong match

#	Patent	Last and first name	Middle name	Soundex code	Street	City
1	5137745	ZUKERMAN RACHAEL	B	Z265500 R240000		NORTHBROOK
2	4764390	ZUKERMAN RACHEL	B	Z265500 R240000	4125 YORKSHIRE	NORTHBROOK
3	5525366	ZUKERMAN RACHEL	B	Z265500 R240000	4125 W YORKSHIRE LA	NORTHBROOK
4	5817355	ZUKERMAN RACHEL	B	Z265500 R240000		NORTHBROOK

#	Zip	State	Partners	Assignee	Patent class	City counter	Assignee counter	Pat-class counter
1		IL	3238287	682323	426	1505	509	30784
2	60062	IL	3238287		426	1505		30784
3	60062	IL	3238287		426	1505		30784
4		IL	3238287	746348	426	1505	2	30784

This example is meant to demonstrate how the CMP works in an "easy" case, in the sense that there is a lot in common between the 4 records. Still, naïve matching would have not worked: even though all records share the same ***last*** name, the first name in record 1 is not identical to the others (Rachael vs. Rachel). Using Soundex, all records are coded the same (Z265500 R240000), and thus belong to the same *p-set*, which is a "rare" one given that there are only 4 records associated with it (less than the cutoff of 16). Consider now records 1 and 2: since the names in them are not identical and the Soundex code for the first name is only 2 digits long, the threshold for a match is 120. The two records share the same city (Northbrook, IL), which we regard as "large" since its patent count is 1,505 (the cutoff for cities is 1,382), thus scoring 100 in that regard. In addition, both records cite patent 3711295 (not shown above), and share the same partner (Harold Zukerman), adding 240 points to the score. Moreover, both records share the same patent class (a "large" one, with 30,874 patents), adding 80 points; the total score is thus 420, which is well above the threshold of 120, and hence they are matched. Matching records 3 and 4 to the first two is even easier, since they share the exact same name as record 2 (hence the threshold is only 100), and all the other criteria are also met: same citation, partner, city and patent class. Notice that the address appears in two of the records (and it is a "strong" criterion"), yet it was not used in the matching, since it is not written ***exactly*** the same way, even though it clearly is the same address.

Example 2:
Deploying transitivity

#	Patent	Last and first name	Middle name	Soundex code	Street	City	State
1	3881067	FLEISCHFRESSER GERALD	H	F421626 G643000		CHICAGO	IL
2	4410982	FLEISCHFRESSER GERALD	H	F421626 G643000		WHEATON	IL
3	5843272	FLEISCHFRESSER GERALD		F421626 G643000		WHEATON	IL

#	Partners	Assignee	Patent Class	City counter	Assignee counter	Pat-class counter
1	677809	234405	379	10763	699	20715
2	677810	234405	370	1852	699	25467
3	677811	377880	156	1852	16634	33598

All three records share the exact same unified "rare" name, and therefore require a score of at least 100 for a match. Records 1 and 2 share the same middle name initial, scoring 100 points, and the same "small" assignee (adding another 100 points), and therefore match with a total score of 200. Records 2 and 3 match as well, since they share the same "small" city, which combined with a low-frequency name scores 100 points, exactly what is required for the match. However, when trying to match records 1 and 3 the score is 0, since except for the name they have nothing in common. Still, the three are matched making use of transitivity: record 1 is matched to record 2, record 2 to 3, and hence 1 to 3 as well, that is, the three are pronounced to be the same inventor and assigned the same ID. Notice that the AMS_i in this case is quite low: (200+0+100)/3=100.

Example 3:
Identifying different inventors with the same name

#	Patent	Last and first name	Middle name	Soundex code	City	State	Country
1	4256297	PINARD PATRICK		P563000 P362000	SEICHES LE LOIR		FR
2	4319745	PINARD PATRICK		P563000 P362000	SEICHES LE LOIR		FR
3	5815811	PINARD PATRICK		P563000 P362000	SANTA CLARA	CA	US
4	6002918	PINARD PATRICK		P563000 P362000	SANTA CLARA	CA	US

#	Citation	Partners	Assignee	Patent Class	City Counter	Assignee Counter	Pat-Class Counter
1			114995	271	2	428	10757
2			114995	271	2	428	10757
3		2222244	676549	455	4627	1230	19592
4		2222244	676549	455	4627	1230	19592

This case exemplifies the advantage of the CMP over naïve matching (i.e. matching just by names): even though all four records bear exactly the same "rare" inventor's name (Pinard Patrick), they correspond to two different inventors, not one. The first two patents are located in the same city, same assignee (Compagnie Internationale pour l'Informatique), and same patent class. Similarly, records 3 and 4 are in the same city (Santa Clara, California), same company (Symbol Technologies, Inc), same co-inventor (Frederic Heiman), and same patent class. By contrast, there is nothing in common between the first two records and the last two. Thus, naïve matching would have grouped the 4 together incurring a Type II error, whereas the CMP correctly identified two different inventors:

#	Name	CMP ID	Total Score
1	PINARD PATRICK	16110559	280
2	PINARD PATRICK	16110559	280
3	PINARD PATRICK	16110561	380
4	PINARD PATRICK	16110561	380