The careers and co-authorship networks of U.S. patent-holders, since 1975.

Ronald Lai
Baker Library
Bloomberg Center 280D
Harvard Business School
Boston, MA 02163
rolai@hbs.edu

Alexander D'Amour
Center for Government and International Studies Knafel K322
Harvard Institute for Quantitative Social Science
Cambridge, MA 02138
adamour@iq.harvard.edu

Lee Fleming
Morgan 485
Harvard Business School
Boston, MA 02163
lfleming@hbs.edu

NOTE: This is a very preliminary paper and only meant to provide the skimpiest of details on the preliminary data that will be made available as soon as we calculate it. We know the data will have many errors and hence strongly suggest that researchers automate any work they invest in applying the data. We would also very much appreciate feedback and help in identifying any problems.

February 13, 2009

We would like to thank the Harvard Business School Department of Research and the NSF for supporting this research. Errors and omissions remain ours (though we ask that you bring them to our attention).

The inventing careers and social networks of U.S. patent-holders, since 1975.

Abstract: We describe disambiguation algorithms that identify individual inventors from the U.S. patent database. The identification enables construction of social networks based on patent co-authorship. We will eventually provide descriptive statistics of individual and collaborative variables and illustrated examples of networks for an individual, an organization, a technology, and a region. The data and code will be publically available for community use and improvement and will enable updating as frequently as new patents are issued.

Introduction

Patent data for the United States became widely available to researchers in the mid 1990s. These data enabled an outpouring of research in the fields of technology and innovation. Though individual researchers had developed their own databases from the raw data, the publication of a dataset from the National Bureau of Economic Research (hereafter referred to as the NBER, see Hall, Trajtenberg, and Jaffe, 2001) enabled a much broader set of researchers to use the patent data. This effort drastically reduced the barriers to entry and made patent data available to a larger community of researchers that lacked the resources, hardware or programming skills to access the data.

The original NBER database included authorship and firm and state level data but did not identify unique inventors over time. This is a non-trivial task because the United States Patent Office (USPTO) did not require consistent and unique identifiers for inventors (for example, the second author of this paper is listed on one of his patents as Lee O. Fleming and the other as Lee Fleming). The aim of this paper and the dataset associated with it is to identify the individual inventors over time and make their social networks (from patent co-authorship) available. The algorithms and code to accomplish will be made public as well in the hope that the community of researchers will take responsibility for future updates and continuous improvement. This paper builds directly on prior efforts by a variety of researchers (Fleming and Juda, 2004; Trajtenberg, Shiff, and Melamed, 2006; Fleming, King, and Juda, 2007; Singh 2007).

Figure 1

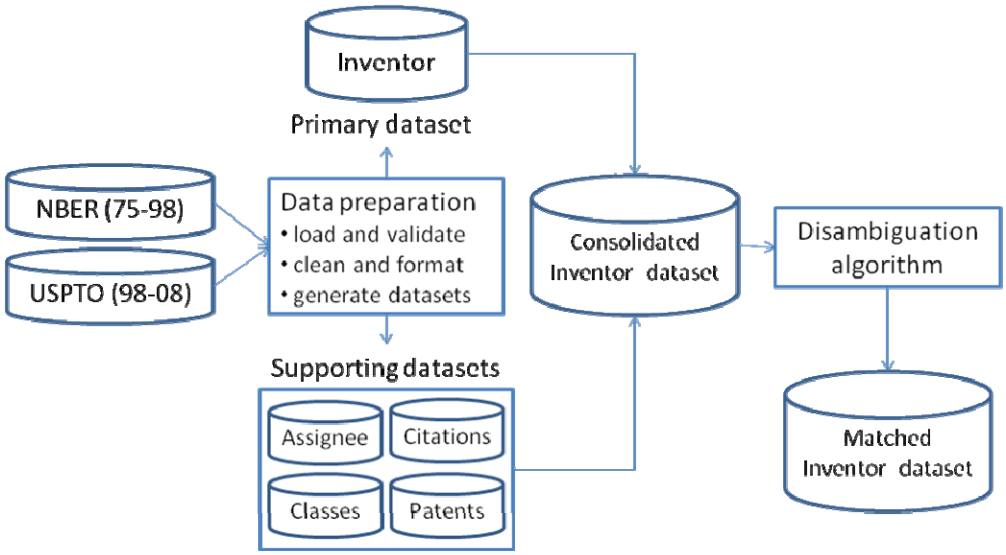
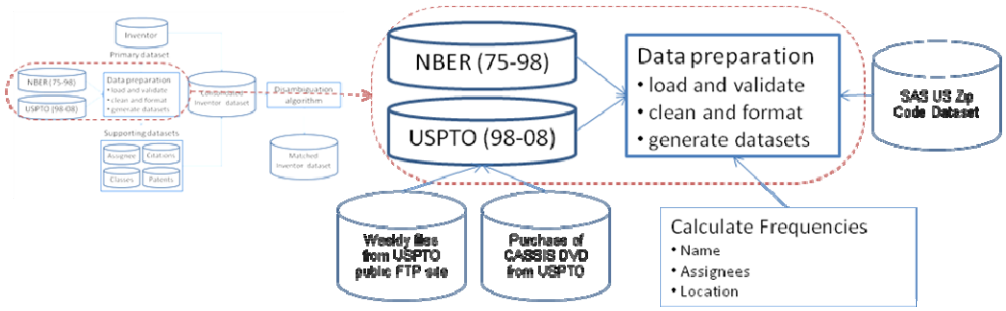


Figure 1 provides a high-level block diagram of the process. Source data come from the NBER database (Hall, Jaffe, and Trajtenberg, 2001) and directly from the US Patent and Trademark Office (USPTO). The data preparation step generates two sets of datasets, first data relevant to the inventor name and second, associated data that will be used to supplement the inventor identification process. These datasets are then fed into the disambiguation algorithm which outputs the matched inventor dataset. The first section ("Overview of dataset creation") provides explanation on how the inventor dataset is created, the second section ("Matching and disambiguation algorithms") details the matching processes, the third section ("Working example: the disambiguation of Matthew T. Marx") provides a working example and the final section ("Potential applications of the data to substantive problems") offers ideas for substantive applications of the data. Appendices include formatting information and a variable dictionary.

Overview of dataset creation

Figure 2



Primary data sources

The final inventor, assignee, citation, patent and classification datasets were built using primary data sources from the USPTO and the NBER. The USPTO makes up-to-date patent data available through their public FTP resource[1] and collaborates with the European and Asian patent offices to provide these weekly updates. Weekly data on newly granted patents are released in a single file which is actually an aggregation of several XML files[2]. In 2006, the USPTO also began providing weekly updates on patent applications. However, since applications are not guaranteed an eventual grant, at this point in time, we have chosen not to incorporate these data. While the USPTO resource provides the data which allows us to keep this dataset current, we extend the dataset to 1975 with the NBER patent database.

The NBER patent database contains patents granted from 1975-1999 and is publicly available[3]. Since the NBER inventor dataset begins in 1975[4], we are utilizing information from 1975

1 USPTO provides a public FTP resource which is updated on a weekly basis and contains raw patent data from late 1996 to the present. ftp://ftp.uspto.gov/pub/patdata/.

2 There have been several observed format changes in the data in 1998, 2001 and 2002. Due to the constant improvement of USPTO information and change in formats, future changes to the data preparation step will surely be needed. The senior author intends to support these changes but in addition, all programming code will be made available so that the community can respond to these issues as well.

3 National Bureau of Economic Research (NBER) has published patent data (Hall, Jaffe, and Trajtenberg, 2001) and provides this as a public resource. More information can be found here: http://www.nber.org/patents/.

onwards. To the best of our knowledge, there is not a comprehensive computer database which contains inventor information before 1975. The NBER database only contains utility patents. However, since utility patents are the dominant patent types (see Table 1 for patent types proportions in 2007), this is not a serious problem.

Table 1

	2007 Patents	
Type	Frequency	Percentage
Design	24,069	13.14%
Utility	157,502	85.99%
Other	1,584	0.87%

To verify our raw dataset, we compared the number of patents granted per year from our dataset to aggregate patent counts from the USPTO. As Table 2 indicates, we have identified small variances between patent counts in our dataset and those from the USPTO and continue efforts to eliminate those variances. We welcome community contributions in this effort.

4 NBER provides limited data from 1963-1999 but only provides inventor data from 1975-1999. Since inventor information is necessary in our disambiguation algorithms, we have only matched inventors to patents granted after 1975. Further information about the inventor dataset can be found at: http://www.nber.org/patents/inventor.txt.

Table 2

Year	Utility Patents			Design Patents			Plant Patents		
	HBS	USPTO	Variance	HBS	USPTO	Variance	HBS	USPTO	Variance
1975	72,000	72,000	0.0%	n.a.	4,282	n.a.	n.a.	150	n.a.
1976	70,226	70,226	0.0%	n.a.	4,564	n.a.	n.a.	176	n.a.
1977	65,269	65,269	0.0%	n.a.	3,929	n.a.	n.a.	173	n.a.
1978	66,102	66,102	0.0%	n.a.	3,862	n.a.	n.a.	186	n.a.
1979	48,854	48,854	0.0%	n.a.	3,119	n.a.	n.a.	131	n.a.
1980	61,819	61,819	0.0%	n.a.	3,949	n.a.	n.a.	117	n.a.
1981	65,771	65,771	0.0%	n.a.	4,745	n.a.	n.a.	183	n.a.
1982	57,888	57,888	0.0%	n.a.	4,944	n.a.	n.a.	173	n.a.
1983	56,860	56,860	0.0%	n.a.	4,563	n.a.	n.a.	197	n.a.
1984	67,200	67,200	0.0%	n.a.	4,938	n.a.	n.a.	212	n.a.
1985	71,661	71,661	0.0%	n.a.	5,066	n.a.	n.a.	242	n.a.
1986	70,860	70,860	0.0%	n.a.	5,518	n.a.	n.a.	224	n.a.
1987	82,952	82,952	0.0%	n.a.	5,959	n.a.	n.a.	229	n.a.
1988	77,924	77,924	0.0%	n.a.	5,679	n.a.	n.a.	425	n.a.
1989	95,537	95,537	0.0%	n.a.	6,092	n.a.	n.a.	587	n.a.
1990	90,364	90,365	0.0%	n.a.	8,024	n.a.	n.a.	318	n.a.
1991	96,513	96,511	0.0%	n.a.	9,569	n.a.	n.a.	353	n.a.
1992	97,444	97,444	0.0%	n.a.	9,269	n.a.	n.a.	321	n.a.
1993	98,342	98,342	0.0%	n.a.	10,630	n.a.	n.a.	442	n.a.
1994	101,676	101,676	0.0%	n.a.	11,095	n.a.	n.a.	499	n.a.
1995	101,419	101,419	0.0%	n.a.	11,712	n.a.	n.a.	387	n.a.
1996	109,645	109,645	0.0%	n.a.	11,410	n.a.	n.a.	362	n.a.
1997	111,984	111,984	0.0%	n.a.	11,414	n.a.	n.a.	394	n.a.
1998	147,574	147,518	0.0%	14,768	14,766	0.0%	561	561	0.0%
1999	153,593	153,485	0.1%	14,735	14,732	0.0%	421	420	0.2%
2000	157,596	157,494	0.1%	17,416	17,413	0.0%	552	548	0.7%
2001	166,065	166,036	0.0%	16,873	16,871	0.0%	584	584	0.0%
2002	167,425	167,331	0.1%	15,451	15,451	0.0%	1,133	1,133	0.0%
2003	169,105	169,023	0.0%	16,576	16,574	0.0%	994	994	0.0%
2004	164,413	164,291	0.1%	15,695	15,695	0.0%	1,016	1,016	0.0%
2005	143,927	143,806	0.1%	12,954	12,950	0.0%	716	716	0.0%
2006	173,922	173,771	0.1%	20,971	20,965	0.0%	1,149	1,149	0.0%
2007	157,502	157,283	0.1%	24,069	24,063	0.0%	1,047	1,047	0.0%
98-07	1,601,122	1,600,038	0.1%	169,508	169,480	0.0%	8,173	8,168	0.1%
Total	3,439,432	3,438,347	0.0%	n.a.	323,812	n.a.	640	14,649	n.a.

Secondary data sources

In addition to the primary data sources, we merged in data from secondary data sources to create better parameters for identifying inventors. These secondary data sources include the USPTO CASSIS dataset and the SAS Institute's zip code file.

When a patent is granted, the USPTO assigns it alphanumeric classes based on the type of technology that it represents. However, due to changes in technology, the USPTO often changes these classifications. The USPTO keeps track of these classification changes in a product called

CASSIS, which is a DVD that is updated every two months[5] and is approximately released on a two month lagged basis. We have used the August 2008 version for our preliminary analysis.

In identifying individual inventors, we found it useful to include geographic parameters as well. To calculate distances between addresses in the patent data, we used an updated US zip code file from the SAS Institute[6] which is easily integrated into the SAS platform. In this version, we used the October 2008 released US zip code dataset. All secondary datasets are released with a lag so we do not expect year end versions until a later date.

Preparing the inventor dataset

To minimize redundancy, rather than generating one large dataset containing all unique combinations of patent information[7], we created several smaller datasets that can be joined together on unique patent and inventor identifiers. These smaller independent datasets are organized by information type. They are assignees, citations, classes, inventors and patents. A detailed description of each of these datasets is provided in Appendix 1.

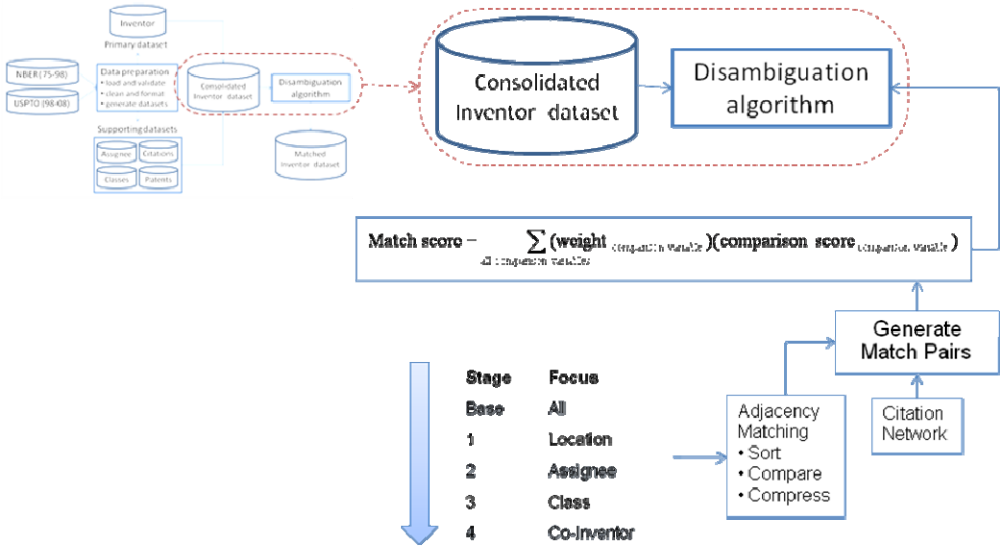
5 'Patents CLASS: Current Classifications of US Patent Grant Publications 1790 to Present' (Code: EIP-2050P-DD): http://www.uspto.gov/web/offices/ac/ido/oeip/catalog/products/pp-o2w-3.htm#classP2050dd

6 SAS provides updated US zip code datasets which can be assessed through its web space: http://support.sas.com/rnd/datavisualization/mapsonline/html/misc.html

7 USPTO patent data contain 60+ fields of information. If we were to restrain our data into one primary dataset, unique permutations of each field would be difficult to manage. For example, many patents contain several inventors (INV), several classifications (CLS) and several citations (CIT). At the very minimum, a dataset would require INV x CLS x CIT records of data. Clearly, this can create an unnecessarily large and clunky dataset.

Matching and disambiguation algorithms

Figure 3



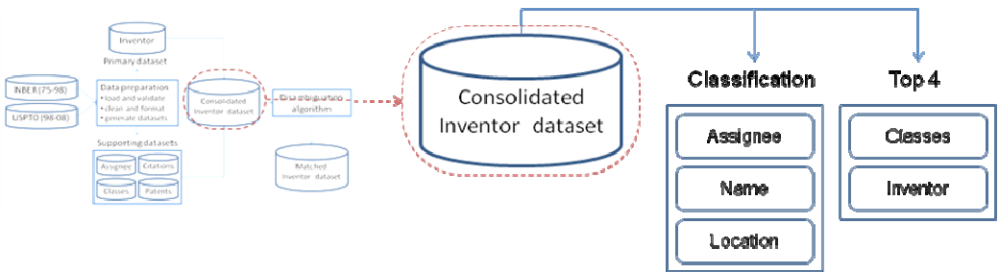
Identifying all the patents of individual inventors requires disambiguating their names. This process presents two main challenges. The first is developing a method for assessing the similarity between two records. No single variable in the available patent data serves as an acceptable indicator or a match. Inventor names, while the most obvious proxy, are unreliable. First names can be an abbreviated form of the full name. Middle names can be represented with a single initial, full middle name, or blank. Foreign last names are susceptible to different phonetic interpretations. A number of names occur so frequently that we risk identifying distinct inventors as the same person. Assignees' names are no better – they often include subsidiary names as substitutes, and components of assignee names are often abbreviated (the assignee owns the patent – in many cases it is left blank, which most often indicates that the inventor owns the patent). A variety of errors (spelling, for example, caused by manual input and automated Optical Character Recognition (OCR) technology) only complicate the process. Previous work in this area (Fleming and Juda, 2004; Fleming, King, and Juda, 2007; Singh 2007; Trajtenberg, Shiff, and Melamed, 2006) approached the problem by matching records with similarities across all available variables. The matching algorithm developed here uses a similar

approach, but employs different approximate and exact matching techniques, along with an optimization of the weights to assign to each comparison.

The second challenge is the sheer size of the dataset. Even with a good method for comparing two records, the size of the dataset makes the cost of doing an exhaustive comparison, an operation that runs in $O(n^2)$ time, prohibitive. Previous research (Trajtenberg 2006 et al) approached the running time issue by partitioning, or blocking[8], the dataset into groups, then performing exhaustive comparisons within those groups. We found that this approach lacked flexibility – once two records are assigned to different blocks, there is no way that they can be matched together – and, depending on the size of blocks, could approach $O(n^2)$ complexity. In sum, either the partitioning function would be too strict, in which case too few records would match, or it would be too lenient, in which case the running time would be too large. Instead, we developed an algorithm we call "adjacency matching" which takes advantage of how the data are arranged, and whose running time is driven by the complexity of sorting a list, O(n log n).

Variable Inclusion and Classification

Figure 4

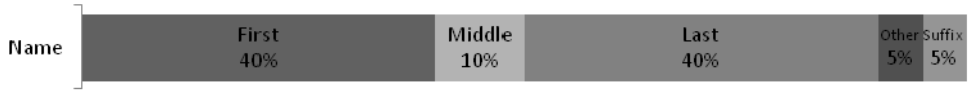


The first question in developing a methodology is which variables to include. We consolidated our primary and secondary data sources into a raw inventor dataset containing over 8 million

8 "Blocking" is a pre-processing step in matching that decreases the number of needed comparisons while not eliminating any potential matches for comparisons. It partitions the matching space into possible and impossible matches with simple and hopefully accurate criteria. While this partitioning is usually exhaustive and mutually exclusive (Herzog, Scheuren, Winkler, 2007, pgs. 123-124), our blocking criteria would need to allow consideration of overlapping sets of potential matches. For example, Matt T. Marx and Matt Marx would be identified as possible matches, as would Matt A. Marx and Matt Marx. These two sets would not be considered together, but Matt Marx would need to be compared *after* blocking to Matt T. *and* Matt A. Marx.

inventor-patent instances (for example, a patent with three inventors would create three inventor-patent instances) with more than 25 variables for each inventor-patent pair. Many of these variables are related, and not all of them are of equal importance in matching two records. We formalized this idea by grouping the variables into categories, and assigning each variable a particular weight within that category. The categories we defined are as follows: name information (variables dealing directly with the inventor's name), assignee information (variables about the organization to which the patent was granted), location information (variables extracted from the address provided by the individual inventor), class information (variables related to the patent's technology class), and coauthor information (the list of coauthors on the patent). Each category's internal weightings are pictured in a series of figures below[9].

Figure 5



While by no means perfect, an inventor's name is still the most obvious feature to start with in disambiguation. In our raw dataset, the author name is split into first name, middle name, last name, suffix, and "other" variables which we chose to weight as above.

Figure 6

When a patent is granted, it is usually assigned to the inventors' employer, which then owns the patent. Since the assignee is often constant over significant stretches of an inventor's patenting career, variables regarding a patent's assignee are important in disambiguating inventors. The raw dataset contains a variable for assignee type (a numerical value corresponding to one of 16

9 The internal weight described here were arbitrarily chosen based on an intuitive understanding of the data. Two more rigorous variable weighting methods are forthcoming, the first of which makes use of training sets and logistic regression to empirically determine suitable weights, and the second of which discards the linear model altogether to allow for non-linear interaction between variables.

categories) in addition to the assignee's name. The assignee name would ideally be enough to identify the firm that holds the patent, but problems arise from misspellings, different forms of the same company's name (IBM and International Business Machines, for example) and the fact that subsidiaries often have completely different names from the parent. Because of these issues, we included the assignee's type in the assignee category, but because of its limited resolution, we gave it a low weight.

Figure 7

Inventors usually include their home address in a patent application, providing another set of variables that are often constant over stretches of their careers. Location information is available in three variables: city, state and country. We chose to include two variables: city and location (loc), where location was defined as the state for domestic patents (Cambridge, MA) and the country for foreign patents (Tokyo, JP). By combining state and country into a single location variable, this category is made robust against potential errors in USPTO data entry where, because the country code is omitted, a state code is misinterpreted as a country code. For example, if we were to compare two records such as “Los Angeles, CA USA” and an incorrectly entered “Los Angeles, CA (Canada)”, city and loc would be equivalent in both pairs resulting in an accurate matching.

In addition to including the name of the inventor's location in the dataset, we also included geographic information for US locations. The SAS Institute provides a regularly updated (last updated August 2008) zip code dataset for US cities which contains additional information including longitudinal and latitude information. This allowed us to see similarity between geographically close but lexically distant cities like “CAMBRIDGE, MA” and “BOSTON, MA”, and tell the difference between cities with the opposite relationship like “STANFORD, CA” and “STAMFORD, CT”. To our knowledge, there is not a longitude and latitude foreign country dataset easily available, so we did not include geographic variables for foreign locations.

Each patent also has technology class and coauthor data which are simple lists of classes and co-authors. These categories provide important information about the inventor's area of expertise and co-authorship network, respectively. We weighted each entry in the list equally, and because of time and space constraints chose to truncate each of these lists at the four most relevant classes (primary only) and co-inventors (firstname and lastname only).

Comparison Strategy

Records were compared using a bottom-up approach. Given two records, we could compare each variable and assign it a similarity score scaled between 0 and 1 where 0 was least similar and 1 was most similar. The similarity score was calculated based on the variable's type and its susceptibility to error. Thus, long character string variables like inventor and assignee names, which are susceptible to typographical errors, were compared using a fuzzy matching technique (see below) which produced a continuous similarity score between 0 and 1, while short character strings and numerical values like assignee type or location code were compared using exact matching which produced a binary similarity score of 0 or 1. If the data fields being compared were lists (for example, lists of co-inventors), each pairwise combination was scored, and the highest similarity score was taken. When available, geographic data were also scored[10].

These individual similarity scores were averaged by category, using the internal weights defined above. Each category score was then adjusted for frequency (see below) and assigned an external weight, and the weighted average of the category similarities was used as the total match score. If the match score was above a certain threshold, we assessed the records as a match. The external weights and threshold were varied depending on the focus of the comparison (see the section titled "Adjacency Matching").

This linear scoring method has several shortcomings, but one is that it assumes that category similarity scores are independent despite there being substantial correlation between category similarities. For example, if a name falls short of a high similarity score, for example, even if it were feasible for the rest of the categories to make up for the shortfall in name similarity, it is

10 An approximation for distance (miles) = $\sqrt{([69.1 * (\text{chg in latitude})]^2 + [53.0 * (\text{chg in longitude})]^2)}$
Probability of distance for longitude and latitude is calculated = min(100% - distance * 0.1%, 0)

very unlikely. To account for this correlation, we added minimum similarity thresholds to each category. If any category (after frequency adjustment) fell short of its internal threshold during record comparison, the process was short-circuited and the records were labeled a non-match. This method had the added benefit of reducing computation time. These thresholds were also varied depending on the focus of the comparison. While this approach might miss a valid match in a particular iteration, the minimum threshold is adjusted for additional iterations of the algorithm. In other runs, the minimum threshold is set lower, and other features are allowed to stand out. This makes the probability of losing a match based on the minimum thresholds fairly small (we discuss plans to optimize the weighting and threshold approaches below).

String Comparison

Since there were spelling errors and inconsistencies in the primary data, we employed an approximate matching technique to lessen the effect of typographical differences when comparing character strings. Within the SAS platform, we implemented the Jaro-Winkler method (Herzog, Scheuren, Winkler) for string[11] comparisons. This approximate matching method compares individual string characters through a statistical approach resulting in a calculated probability that two distinct strings are equivalent (for a detailed description of the Jaro-Winkler method, see Chapter 13 of Herzog, Scheuren, and Winkler, 2007). Presented in the table below are a series of example string pairs and calculated probabilities.

Table 3

Original	Comparison	Jaro-Winkler%
MATTHEW	MATTHEW	100.0%
MATTHEW	HEWMATT	95.2%
MATTHEW	MATEW	93.3%
MATTHEW	MATT	91.4%
MATTHEW	MTATWEH	91.4%
MATTHEW	M	74.3%
MATTHEW	TALIN	56.2%
MATTHEW	XYZ	0.0%

A few disadvantages exist in the Jaro-Winkler method. The method is fairly generous in assessing match probabilities. As an illustration, two strings which contain clearly conflicting

11 A string is defined as a variable type which contains a sequence of alphanumeric/symbolic characters. A character is defined as a single alphanumeric/symbolic within a string construct.

information such as “SPEECHWORKS INTERNATIONAL INC” and “TELLME NETWORKS INC” earn a 74.7% probability match score based on the underlying characters. We corrected this problem by defining a high percentage match thresholds. Another disadvantage in utilizing the Jaro-Winkler method was that it is not included with the SAS package. This required the development of a script which had noticeable runtime inefficiencies.

Other methods of sting matching exist including phonetic algorithms such as SoundEx, which have been used in past research (Trajtenberg, Shiff, and Melamed 2006). Although SoundEx is built into the SAS platform and is easy to implement, we found that it resulted in a higher level of improper matches. Although there are extensions of the SoundEx algorithm that improve its strictness and ability to differentiate names from other languages, we chose not to implement them here, mainly due to run-time issues.

Frequency Adjustment

Before combining the category similarity scores into a final match score, we adjusted them to reflect how rarely the entries in that category occurred. Similarity between two sets of rare entries is more indicative of a match than the same similarity score between sets of common entries. For example, there exists a high probability that several inventors would have a very common name such as “DAVID SMITH” and therefore a higher level of scrutiny is necessary before declaring a match. On the contrary, unique names should contribute positively to a match. To determine the frequency of the names, we chose to look at frequency within the dataset, as opposed to relying upon external data such as census data histograms of name frequency. We chose this method because it does not depend on the assumption that internal and external populations are similar (though for a more exhaustive discussion of the tradeoffs, see Trajtenberg, Shiff, and Melamed 2006). First, we isolated relevant variables such as assignee name and geographical location and determined their frequency. We then calculated for each item a frequency percentile, which we defined as the percentage of items that occurred with higher frequency than the current item – thus the most common names would have a frequency percentile close to 0%, and the least common would have a frequency percentile close to 100%.

We then decided that the frequency of an entry could augment or diminish a category score by 5%, so we defined the frequency adjustment as:

$$FA = 95\% + 10\% * (\text{frequency percentile})$$

The frequency adjustments thus ranged from 95% (for the most common entries) to 105% (for the least common entries). Category similarity scores were multiplied by their frequency adjustment before being averaged together to generate the total match score. We have detailed the most frequent data observations for assignees, location and name (lastname, firstname only) in the subsequent table.

Table 4

Ranking	Assignee Name	Freq	Location	Freq	Name	Freq
1	[BLANK]	326,014	Tokyo, JP	260,343	Silverbrook, Kia	2,407
2	International Business Machines Corporation	51,655	Yokohama, JP	119,203	Yamazaki, Shunpei	2,038
3	Canon Kabushiki Kaisha	36,670	Kanagawa, JP	104,095	Kim, Young	1,794
4	Hhitachi LTD	31,200	Kawasaki, JP	80,924	Lee, Sang	1,543
5	General Electric Company	28,425	San Jose, CA US	71,163	Smith, David	1,201
6	Matsushita Electric Industrial Co LTD	26,866	Osaka, JP	63,396	Kim, Jong	1,141
7	Sony Corporation	25,201	Austin, TX US	49,244	Smith, Robert	1,131
8	Samsung Electronics Co LTD	24,450	Houston, TX US	43,392	Kim, Jin	1,072
9	NEC Corporation	22,491	Seoul, KR	39,179	Kim, Sung	1,024
10	Mitsubishi Denki Kabushiki Kaisha	21,042	San Diego, CA US	38,824	Lee, Jae	1,009

Adjacency Matching

Having determined *how* to make effective comparisons, we moved to the question of *which* comparisons to make. Previous disambiguation approaches assumed that an exhaustive comparison would be necessary to generate an adequate matching of inventor records. Resigned to this approach, researchers instead focused on how to pre-partition the dataset into not inconsistent sets within which an exhaustive comparison could be run to keep the processing tractable. We chose not to take the exhaustive search for granted and instead considered leveraging a property of matching which previous authors had considered an afterthought: transitivity. In previous studies, authors determined that to create consistent matchings, if record A is similar to record B, and record B to record C, then records A and C should be matched *regardless of their similarity*. Our algorithm takes advantage of this fact, and operates on the principle that if comparisons are performed in the right order, only a linear number of comparisons should be necessary to create a correct matching. Adjacency matching tries to

approximate this "right order" of comparisons by sorting the list in such a way that each entry is adjacent to an entry that it is likely to match, and formalizes transitivity by compressing similar records into a single entry.

Adjacency matching is an iterative process involving three steps per iteration, each of which is an operation performed on the entire list. Each iteration focuses on a particular category of variable, for example name, assignee, or location. Once the focus is chosen, the steps are to sort, compare, and compress. The sorting step arranges the data in such a way that adjacent records are likely to match given the category of focus. The comparison step compares each record to the previous one using a comparison operation determined by the category of focus, and generates an overall similarity score. The compression step combines all records whose similarity scores are past a certain threshold, retaining all unique information. Assuming that the sorting and comparison operations are approximately coherent in each iteration[12], this is an efficient heuristic to a clustering algorithm and with some transformations and adjustments to maintain consistency could be shown to be isomorphic to a clustering approach[13]. After all iterations are completed, each remaining record is considered to represent a unique inventor.

Adjacency matching is a generic algorithm whose components can be mixed and matched to satisfy whatever conditions the researcher desires. One can run any number of iterations with any number of focus categories until acceptable convergence[14] is reached. One could also use

12 The criterion for true coherence is that for each record, as long as there exist other records in the dataset that are similar enough to be combined with the current record, that at least one of those records be placed adjacent to it. The comparison and sorting operations used to create this dataset were not formally checked to satisfy this coherence criterion, nor do we currently have a method for constructing such a proof. However, the results from the dataset that were manually confirmed indicate that the combinations of sort and comparison operations used here were a good enough approximation to truly coherent operations.

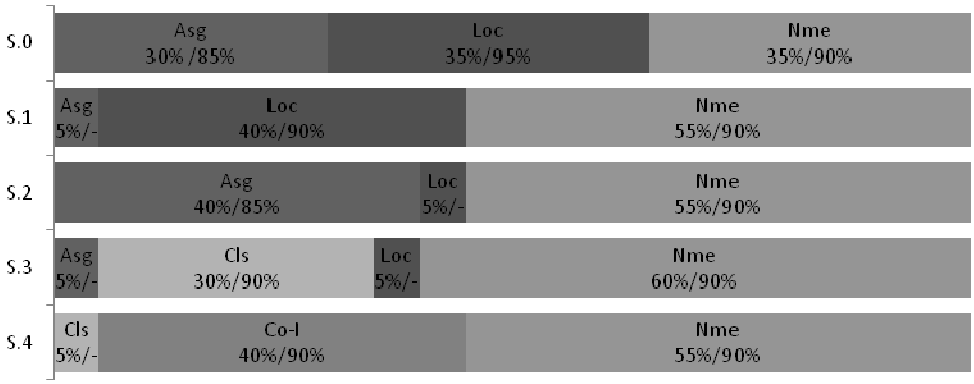
13 The linear scoring method and the threshold criterion for combining two records define an acceptance region around each data point that is a stretched (possibly truncated) simplex when similarity scores are appropriately transformed into distances. The version of the adjacency matching algorithm run to produce the current dataset was not calibrated to be consistent in any transformation to a clustering approach, and is thus presented as a heuristic. The conditions that the scoring functions need to meet to implement a clustering routine using an adjacency matching algorithm are a topic of future research.

14 Convergence in this sense means that the number of records lost of compression in each iteration reaches an acceptably low level. Be advised that convergence in one iteration of the algorithm does not guarantee

different sort, comparison, or scoring operations from the ones presented here. Indeed, because the parameter space of this algorithm is the function space of sort, compare, and scoring operations, there is ample room for improving patent matching while still implementing an adjacency matching framework.

In this implementation, we grouped the adjacency matching iterations into stages, where iterations within a stage maintained the same focus but used several different sorting techniques to reduce the chance of misspellings keeping the proper comparisons from occurring. In some cases we used simple string manipulations such as utilizing only the first few characters or reversing the sequence of characters when sorting the data. The following chart illustrates the comparison method used in each stage (indicated as "S") that we ran in this implementation of the algorithm. Each stage had a different focus, which is reflected in the variable classifications (Assigness as Asg, Classes as Cls, Co-Inventors as Co-I, Location as Loc, and Name as Nme), category weights (first percentage) and the internal category thresholds (second percentage). The assumption weights for each stage were chosen arbitrarily to reflect our intuition about the data and were adjusted based on results from test runs.

Figure 8



convergence in another that does not have the same focus category. Conditions for total convergence given the researcher's choice of sort, comparison, and scoring functions is an area of future research.

Working example: the disambiguation of Matthew T. Marx

In this section, we illustrate the disambiguation algorithm with the patents of Matt Marx, a researcher who has developed and published from previous versions of the database (Marx, Strumsky, and Fleming, forthcoming). We chose Marx as an example because we can personally verify the accuracy of his data and because he has several granted patents which are owned by three different firms, has moved across the country from Massachusetts to California, and has variations within his name. We have confirmed Matt Marx's list of patents personally, a list which also can be found through USPTO resources[15].

Base Stage

To match the correct Matt Marx, we start with the base stage which incorporates three of our defined classifications: Name, Location and Assignee. Patent number is included to assist with identification.

Table 5

Patent	Name	Location	Assignee
5995928	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6173266	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6606598	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6941268	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7140004	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7143039	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7308408	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION
7321856	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

We begin by sorting our Matt Marx subset by name and then application date. This creates natural pairings represented in Table 6, for example, patent 5995928 will then be compared to 6173266, 6173266 to 6606598 and so on.

15 Matt Marx's Patents can be found with use of the following URL: http://patft.uspto.gov/netahtml/PTO/search-adv.htm in combination with the search query: IN/"Marx, Matthew T" OR IN/"Marx, Matthew Talin"

Table 6

Patent	Name	Location	Assignee
5995928	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6173266	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6606598	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6941268	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7140004	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7143039	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7308408	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION
7321856	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

As mentioned above, the base stage category weights and category thresholds are: Name (35% / 90%), Location (35% / 95%), Assignee (30% / 85%) and an overall threshold of 90%.

The comparison process for two of the above pairings is shown below.

Table 7

Patent	Name	Location	Assignee
5995928	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6173266	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
	Match After Freq: 99.7%	Match After Freq: 103.3%	Match After Freq: 103.1%
	Threshold: > 90%	Threshold: > 95%	Threshold: > 85%
	Weight: 35%	Weight: 35%	Weight: 30%
	Scenario Match: 102%		
	Threshold: 102% > 90%	<< Match!	

Patent	Name	Location	Assignee
6606598	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6941268	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
	Match After Freq: 97.0%	Match After Freq: 49.4%	Match After Freq: 79.6%
	Threshold: > 90%	Threshold: < 95%	Threshold: < 85%
	Weight: 35%	Weight: 35%	Weight: 30%
	Classification Threshold Fail!	<< Non-Match!	

Each of the comparisons performed in this stage is a string comparison. The "Match after Freq" field shows the weighted Jaro-Winkler similarity average of the first, middle, and last names after frequency adjustment. The adjustment was calculated using the frequency percentile of Matthew T. Marx (47.2%) and frequency adjustment formula yields 99.7%. Thus, in the first example, despite the names yielding an average Jaro-Winkler score of 100%, the score is scaled back to 99.7%. In the second example, the first and last names are identical, but the middle name does not match exactly. Since "T" could represent another middle name, this comparison

indicates that we have less confidence in the records' matching. The Jaro-Winkler similarity between “T” and “TALIN” is 76.0% so the weighted average similarity of the name category is 97.5%, which is adjusted down for frequency to 97.2%.

After replicating this analysis over the other two categories, we see that the first comparison pair is successful whereas the second comparison is not successful. In particular, the comparison failed to reach the internal Location and Assignee thresholds, so the comparison was short-circuited. This does not mean that we have ruled out the second pairing as a potential match; patents 6606598 and 6941268 may be matched together in a later stage. After the base stage, the records are organized and grouped as follows:

Table 8

Inventor#	Name	Location	Assignee
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

Each Matt Marx unique inventor record is separated with highlighting and a unique inventor number. The inventor number is flexible and relies on three pieces of information: the patent number, inventor sequence, and patent type. Inventor 59959280021 is generated as follows. 5995928 is the patent number, 002 is the inventor sequence[16] and finally the patent type “U” (or Utility) is given a code of 1[17]. Therefore 5995928|002|1 is a unique identifier for Matt Marx. With this logic, the second record would be given an inventor number of 61732660011. When records are matched, we assign both records the smaller of the two inventor numbers, so that

16 Matthew Marx is the secondary inventor for patent 5995928.

17 Patent type code translation: Utility: 1, Design: 2, SIR: 5, X-Series: 6, Plant: 7, Re-issue: 8. Codes 3, 4 are types “A” and “T”. These codes are previous codes and are no longer utilized by the USPTO.

each inventor is identified by the information contained in his or her first record in the dataset[18]. When the matching is completed, this inventor number serves as a unique identifier.

Matches are then compressed by inventor number reducing the number of records and increasing the efficiency of future iterations. This compression also allows a higher chance for the comparison of two originally non-adjacent cells, formalizing the idea of transitivity. When records are compressed, all of the fields, including those not relevant to the current stage's comparison, are merged and redundant information is eliminated. In the table below, we show how the class and co-inventor fields, which were not used in the base stage, are merged:

Table 9

Inventor#	Name	Class	Co Inventors
59959280021	MARX, MATTHEW T	704	MARK HOLTHOUSE, JOHN NGUYEN
59959280021	MARX, MATTHEW T	~~704~~	JERRY CARTER, MICHAEL PHILLIPS, ~~MARK HOLTHOUSE~~
59959280021	MARX, MATTHEW T	~~704~~	~~JOHN NGUYEN~~
69412680041	MARX, MATTHEW TALIN	704, 379	LISA STIFELMAN, HADI PARTOVI, HALEH PARTOVI, DAVID ALPERT
69412680041	MARX, MATTHEW TALIN	717, ~~379~~	JEFF KUNINS, ~~HADI PARTOVI~~, BRANDON PORTER
69412680041	MARX, MATTHEW TALIN	~~704~~, 715	~~BRANDON PORTER~~, ~~LISA STIFELMAN~~, MICHAEL BODELL
73084080051	MARX, MATTHEW TALIN	704, 379	BRANDON PORTER, LISA STIFELMAN, MICHAEL BODELL
73084080051	MARX, MATTHEW TALIN	~~704~~	~~LISA STIFELMAN~~, HADI PARTOVI, HALEH PARTOVI, DAVID ALPERT

After compression, the records appear as follows:

Table 10

Inventor#	Name	Location	Assignee
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

18 Utility patents are far more common than non utility patents. Therefore, by default utility patent number will be far greater than all other patents. If a previous all utility granted inventor is granted a new non-utility patent, all patents will revert to the newer inventor number causing the necessity for a future research to "look up" the inventor. For this reason, there is consideration in implementing grant dates into the unique number. Although this is not estimated to occur very frequently, it could cause results to change as the data changes. Other extensions include adding new inventor information which will cause two previous "unique" inventors to merge.

Stage 1

Matt Marx has now been compressed from eight to three unique inventor records. We move on to stage 1 which focuses on location: Assignee (5%/0%), Location (40%, 90%) and Name (55%, 90%), and overall threshold of 92.5%. The data are sorted on location, name and application date, then compared as shown:

Table 11

Inventor#	Name	Location	Assignee
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

The stage results in a preliminary match for Inventors 69412680041 and 73084080051 and is detailed below.

Table 12

Inventor#	Name	Location	Assignee
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
	Match After Freq: 97.0%	Match After Freq: 49.4%	Match After Freq: 79.6%
	Threshold: > 90%	Threshold: < 90%	Threshold: > 0%
	Weight: 55%	Weight: 40%	Weight: 5%
	Classification Threshold Fail!	<< Non-Match!	

Inventor#	Name	Location	Assignee
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION
	Match After Freq: 99.7%	Match After Freq: 96.8%	Match After Freq: 68.4%
	Threshold: > 90%	Threshold: > 95%	Threshold: > 0%
	Weight: 55%	Weight: 40%	Weight: 5%
	Scenario Match: 95.3%		
	Threshold: 95.3% > 92.5%	<< Match!	

Due to the corroborating data from the second comparison, we are able to further compress our inventor records.

Table 13

Inventor#	Name	Location	Assignee
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION TELLME NETWORKS INC

As previously mentioned, each iteration contains its own sorting operation. To demonstrate the importance of sort sequence, had we sorted this stage's records using a sorting operation from the

base stage (sorting on name followed by assignee, location and application date), we would produce the following sequence which would not only be less ideal in the adjacency matching algorithm but also would produce non matched comparisons.

Table 14

Inventor#	Name	Location	Assignee
69412680041	MARX, MATTHEW T	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
59959280021	MARX, MATTHEW TALIN	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
73084080051	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

Stages 2 and 3

Stage 2 focuses on assignee with weights/thresholds Assignee (40%/85%), Location (5%, 0%), and Name (55%, 90%). We compare a record with one assignee against a second record with multiple assignees. The algorithm considers all possible combinations, in this case, "SPEECHWORKS INTERNATIONAL INC" against "MICROSOFT CORPORATION" and "SPEECHWORKS INTERNATIONAL INC" against "TELL ME NETWORKS INC". The highest match percentage is retained, and the frequency adjustment of the more common assignee is applied. Thus, in the Microsoft comparison, Microsoft's adjustment is applied. Unfortunately, the highest frequency-adjusted similarity score is 75.4%, which fails the internal assignee category threshold of 85%, so stage 2 produces no matches.

Stage 3

Finally, on stage 3, we achieve a match. This stage focuses on patent class, and uses the following weights/thresholds: Assignee (5%/0%), Location (5%/0%), Class (30%/90%) and Name (60%/90%). The overall threshold has been defined as 92.5%

Table 15

Inventor#	Name	Location	Assignee	Classes
59959280021	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC	704
69412680041	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION TELLME NETWORKS INC	704, 379, 717, 715
	Match After Freq: 97.0%	Match After Freq: 49.4%	Match After Freq: 75.4%	Match: 100%
	Threshold: > 90%	Threshold: > 0%	Threshold: > 0%	Threshold: > 90%
	Weight: 60%	Weight: 5%	Weight: 5%	Weight: 30%
	Scenario Match: 94.5%			
	Threshold: 94.7% > 92.5%	<< Match!		

With classes, just as with assignees, when there are multiple entries associated with each record, there is an exhaustive comparison of the lists, where the highest similarity score is taken. In this case class receives a match of 100% since 704 is contained in both lists. Since the frequency class entries is fairly uniformly distributed in the dataset, we chose not do apply a frequency adjustment to class comparisons.

As of this point, we have provided all proper instances of Matt Marx with a unique inventor number of 59959280021. The compressed record now contains several pieces of information.

Table 16

Inventor#	Name	Location	Assignee	
59959280021	MARX, MATTHEW TALIN	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC	
		MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION TELLME NETWORKS INC	
		Class	**Co Inventors**	
		704, 379, 717, 715	MARK HOLTHOUSE, JOHN NGUYEN, JERRY CARTER, MICHAEL PHILLIPS, LISA STIFELMAN, HADI PARTOVI, HALEH PARTOVI, DAVID ALPERT, JEFF KUNINS, BRANDON PORTER, MICHAEL BODELL	

Stage 4

Because the records in the example were already consolidated into one, stage 4 is not necessary here. However, we see clear overlap among co-inventors between several of the patents in table 9. The first two records of Matt Marx show an inventor who has worked on a similar technology class (704) with the same inventor (John Nguyen). The comparison further validates the earlier comparisons (based on name, location and assignee). In addition, although inventors 69412680041 and 73084080051 have differing assignees, namely Microsoft and Tellme Networks Inc, both contain similar patent classes (704, 379) and have many of the same inventors (incl. Matthew Marx, Brandon Porter, Lisa Stifelman, and Michael Bodell). With such consistency, there is a high probability that the records would have been matched together in stage 4 had previous stages missed them.

Observations from the adjacency matching stages hint at information besides author identities that can be gleaned from a well-disambiguated inventor dataset. For example, the overlap of coinventors and technology classes between Matt Marx's Tellme and Microsoft patents suggest a

number of underlying transitions that may have occurred. Hypotheses might include Microsoft acquiring TellMe, or an entire team of inventors leaving Tellme for Microsoft, or the possibility of Marx collaborating with colleagues across corporate boundaries.

Avoiding false positive matches

The dataset we presented in the last section was small with an initial eight records. If we were to expand the data to contain all "MATTHEW MARX" within the data, we would add five records.

Table 17

Patent	Name	Location	Assignee
6492383	MARX, MATTHEW ARNOLD	WATERFORD, CT 06385	PFIZER INC
6964961	MARX, MATTHEW ARNOLD	WATERFORD, CT 06385	PFIZER INC
6995171	MARX, MATTHEW A	WATERFORD, CT 06385	AGOURON PHARMACEUTICALS INC
7271262	MARX, MATTHEW A	WATERFORD, CT 06385	PFIZER INC
7285077	MARX, MATTHEW	PLAINVIEW, NY 11803	[BLANK]
5995928	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6173266	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6606598	MARX, MATTHEW T	EVERETT, MA 02149	SPEECHWORKS INTERNATIONAL INC
6941268	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7140004	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7143039	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	TELLME NETWORKS INC
7308408	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION
7321856	MARX, MATTHEW TALIN	MOUNTAIN VIEW, CA 94035	MICROSOFT CORPORATION

Running the stages we previously highlighted, we can generate three unique inventors: "MATTHEW TALIN MARX", "MATTHEW MARX" and "MATTHEW ARNOLD MARX".

Table 18

Inventor#	Name	Location	Assignee	Class
72850770011	MARX, MATTHEW	PLAINVIEW, NY 11803	[BLANK]	482
64923830031	MARX, MATTHEW ARNOLD	WATERFORD, CT 06385	AGOURON PHARMACEUTICALS INC PFIZER INC	514, 544, 546
59959280021	MARX, MATTHEW TALIN	EVERETT, MA 02149 MOUNTAIN VIEW, CA 94035	SPEECHWORKS INTERNATIONAL INC MICROSOFT CORPORATION TELLME NETWORKS INC	704, 379, 717, 715

Given that the middle names differ incompatibly, "MATTHEW ARNOLD MARX" and "MATTHEW TALIN MARX" appear to be very different individuals. "MATTHEW ARNOLD MARX" is probably an inventor within the pharmaceutical industry given his assignees and the patent classes[19]. A trip from Waterford, CT to Everett, MA would require over 100 miles of

19 Definition for patent classifications can be determined through the following website: http://www.uspto.gov/go/classification/

travel thereby reducing the probability that it is one inventor that simply moved addresses. A similar logic exists when comparing "MATTHEW TALIN MARX" against "MATTHEW MARX", although they should not be split into two different people solely because one was missing a middle name field.

Table 19

Inventor#	Name	Co-Inventor	
72850770011	MARX, MATTHEW	[BLANK]	
64923830031	MARX, MATTHEW ARNOLD	MICHAEL MUNCHHOF, SUSAN SOBOLOV JAYNES, MICHAEL LUZZIO, BINGWEI YANG, CHRISTOPHER AUTRY, SUSAN LA GRECA, JOEL ARCARI, JINSHAN CHEN	
59959280021	MARX, MATTHEW TALIN	MARK HOLTHOUSE, JOHN NGUYEN, JERRY CARTER, MICHAEL PHILLIPS, LISA STIFELMAN, HADI PARTOVI, HALEH PARTOVI, DAVID ALPERT, JEFF KUNINS, BRANDON PORTER, MICHAEL BODELL	

Looking at the co-inventor data, no overlap exists for "MATTHEW ARNOLD MARX" and "MATTHEW TALIN MARX". As a whole, we can see why, even after all five stages of comparisons, these records remained separate.

Final Remarks on matching

The adjacency matching process is and shall always remain imperfect. A major issue is that adjacency matching will miss comparisons between records that aren't adjacent, and work beyond modifying the sorting operation may be necessary to capture comparisons that we missed. This is especially worrisome for variations in names that cannot be captured by fuzzy matching, for example Robert and Bob, or differences in names between parent companies and subsidiaries. One possible solution is to pre-process the data by building a hash table which maps disparate names that are known to point to the same entity to a standardized form. However, this table will in many cases need to be updated manually. A community maintained hash table that compiles the results of previous manual work would be an invaluable resource, and a proposal for such a community maintained dataset is forthcoming. Such a table would substantially raise the upper bound on accuracy that automatic matching algorithms could achieve.

In other cases, adjacency matching may lump together two inventors whose records in the dataset are too similar. This occurs quite often for East Asian inventors because East Asian names are often phoneticized similarly in English, and because patenting is often concentrated in particular industries that are dominated by large conglomerates. This is only exacerbated by that fact that the resolution of our geographical data for non-US locations is not as good. Future work will aim to quantify the magnitude of these problems and gather additional data to make lumping more unlikely.

Another improvement which we implemented for this dataset was to match the dataset to the best of our ability using adjacency matching, and then to use citation data to discover new comparisons. To try to recover comparisons that adjacency matching may have missed, we finalized our dataset by exploring the network created by patent citations. Patents are first identified only by primary inventors, which are then compared to the inventors of the cited patents. If there is a strong match in the inventor's name, a comparison for assignee and location made to provide whether a strong indication exists that the inventor has made a self citation. The usage of citation data is not transitive and currently measures only direct relationships. Another technique to form comparison pairs of misspelled information is to utilize linked networks of information. For example, we have utilized the natural patent citation network. Since it is well known that authors oftentimes cite their previous work to develop current or future work, citations provide useful network of information. Once validated as a self-citation, we are able to easily match two records which may not have been classified as a pair in the original adjacency comparison algorithm. Since no single matching algorithm appears to be the silver bullet, layering multiple algorithms on top of one another seems to be an effective way to improve accuracy. Computation currently requires approximately 24 hours on the school's research server[20].

20 http://intranet.hbs.edu/dept/research/unix/specifications.html
System hardware specifications:
- CPUs: 8 64-bit UltraSPARC IV @ 1.35 GHz
- Threads-per-CPU capability: 2 (each CPU is dual-core)
- Total simultaneous processes active at once: 16 (2 per CPU)
- CPU Cache: 16 MB cache per CPU
- Memory: 32 GB RAM

We also intend to implement a learning algorithm approach that has been used to disambiguate the Medline database (Torvik et al. 2005). The approach takes very likely matches and computes a non-parametric distribution to describe those matches. It then takes randomly chosen matches which are very unlikely. These distributions are then used to compare and consider the rest of the dataset. As an intermediate step in this direction, we will be implementing the idea with a simple logit estimation and then using the coefficients of this estimation to weight our linear approach accordingly.

- Networking: 1 Gbit Ethernet
- Solaris 10 64-bit UNIX operating system

Solaris fair-share scheduler to balance processing: each user shares CPU capacity if system becomes overtaxed, but has full access to all CPU resources if run level is below 100%.

Potential applications of the data to substantive problems

We would like to spend some time sketching how these data might be used substantively. We do this in our area of inquiry, namely creativity, collaboration, and inventor mobility, but our hope is that the larger community will apply these data in completely unforeseen and novel ways.

Individual level of analysis research that will be enabled by the data:

At the individual level, the primary benefit of a large and longitudinal database of collaborative relationships will be to complement the mainly survey-based network research that has occurred to date. Much social network research relies upon questionnaires and field work. While field methods enable rich and real-time data collection, they are limited in other ways. For example, they are expensive to administer repeatedly, must inevitably bound their networks, usually sample without an understanding of the underlying distribution of nodes, and often rely upon self-report. Furthermore, because data collection is so expensive, field studies tend to analyze all collected data. This leads to spurious correlation in statistical models, due to a lack of independence between adjacent nodes. Archival databases can complement field survey networks, because they can observe an individual over a lifetime, do not require bounding (for example, the patent data have revealed one connected component with over 100,000 authors, see Fleming, Mingo, and Chen 2007), and can more easily measure phenomena across large and indirect networks (for example, the generation and subsequent diffusion of ideas). Furthermore, if enough data exist, they can be sampled randomly, and hence avoid network autocorrelation problems.

The database will allow the exploration of many questions in creativity, inventor mobility, and the diffusion of ideas. For example, what portion of the advantage of collaboration is the enhanced opportunity to generate new ideas vs. the benefit that comes from easier diffusion of the ideas that are generated? In other words, does collaboration result in higher quality ideas, or does collaboration simply make it more likely that those ideas will be transmitted to and used by others, and be perceived as higher quality? What is the creative benefit from spanning a technological or organizational boundary, as opposed to spanning social space? Social and technological boundaries correlate, but they are distinct constructs, with different effects and

interactions (Fleming and Waguespack, 2007). What is the optimal collaborative structure at different points in an inventive career? In other words, should a recently-graduated inventor collaborate differently than an experience inventor? Do collaboration networks vary by gender or ethnicity and is there a creative benefit to such differences? Finally, how does knowledge diffuse from scientific networks to technological networks (this will require merging of the patent and paper databases).

In addition to basic correlations between social network position and a variety of outcomes, the large database will enable matching and other approaches for developing appropriate controls. The database will also greatly facilitate field work and the selection of interviews to complement the archival data (Fleming, King, and Juda, 2007). The influence of social networks on inventor mobility can be studied across 50 years (thus enabling empirical refinement of Granovetter, 1973), and the influence of direct and indirect networks upon the diffusion of knowledge can be traced across individuals, organizations, regions, and technology domains. Perhaps most importantly, the archival nature of this database would enable the study of social networks over the entire careers of inventors, and how those networks enable creativity, mobility, and the co-evolution of science and technology with careers. Ultimately, if ethical concerns about the identification of individuals can be resolved (perhaps through the merging of datasets behind "enclave" firewalls), these data could be merged with individual data on births, deaths, and family and employment events. This would enable stronger causal inference, based on natural experiments (for example, see Azoulay, Zivin and Wang, 2007).

Organizational level of analysis research that will be enabled by the database:

The availability of the NBER database has enabled a great deal of research on firms and patenting and some of this research has considered social networks. For the most part, however, this research has aggregated patent data to the organizational level, for example, citations or counts. This proposal will enable organizational level analysis to use the underlying patent co-authorships in a much more detailed and flexible way. For example, density measures of collaboration within the firm could be calculated (number of co-authorships divided by number

of patents). More interestingly, the proposal would enable the study of the typologies and topologies of collaboration within and across organizations.

Figures 9 and 10 provide an example of the radical differences in collaborative structure across firms. They illustrate two firms within the carburetion industry, circa 1990, when the industry was undergoing a radical technology shift from mechanical to electronic carburetion. The figures offer stark contrasts into how the firms' inventors collaborated. Merged data from outside the patent record would see to connect these graph types with technological, market, and strategic outcomes, possibly using field interviews, product performance archives, and statistical methods. With the proposal, such research questions could be quickly and efficiently studied in a variety of industries, both visually and statistically.

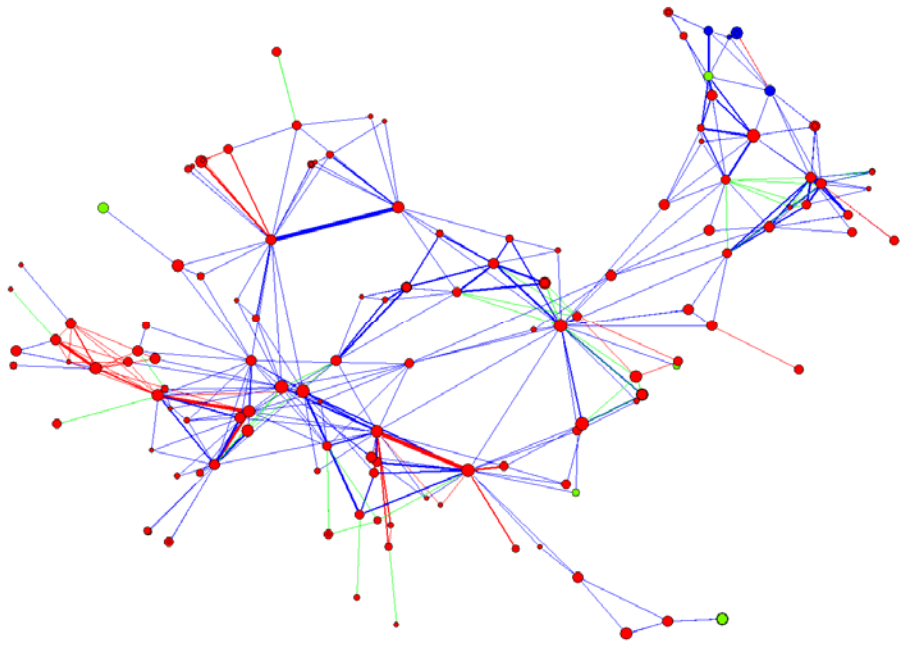


Figure 9: Bosch carburetor patents, circa 1980 (unpublished, developed with Dan Snow and Venkat Kuppuswamy). Note the difference with Figure 10, in that Bosch is much more collaborative. Nodes represent inventors and node size corresponds to the number of patents. Black nodes represent inventors who work in physical technologies, dark grey nodes represent electronic technologies, and light grey nodes represent inventors in both technologies. Tie width corresponds to the number of co-authored patents. Light grey ties represent later ties, black ties earlier ties, and dark gray ties intermediate between early and late.

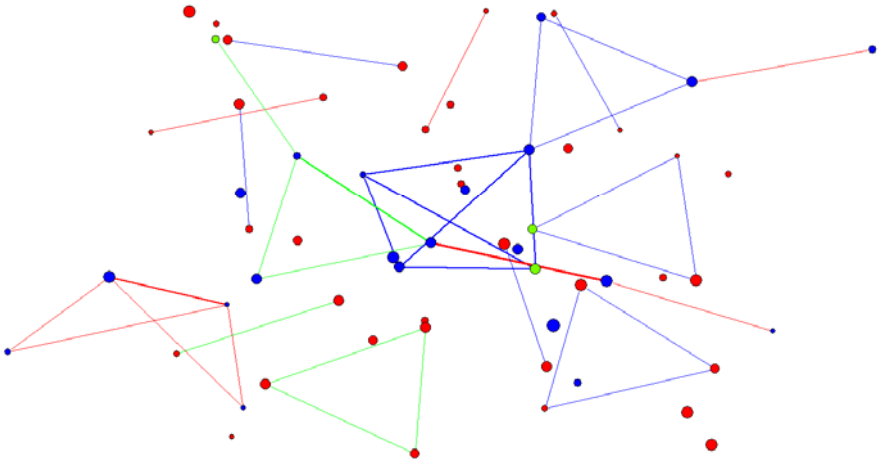


Figure 10: Ford carburetor patents, circa 1980 (unpublished, developed with Dan Snow and Venkat Kuppuswamy). Ford inventors are much more isolated and less collaborative than Bosch inventors illustrated in Figure 9.

Regional level of analysis research that will be enabled by the database:

Social networks have long been argued to be an integral part of regional innovative dynamics (Marshall, 1919; Saxenian, 1994; Owen-Smith and Powell 2004). Complete archival data to study such arguments are difficult to compile, however, such that wide and long panel datasets remain rare (for example, see Fleming, King and Juda, 2007). Due to zip code data associated with the first inventor (early data are spottier), many patents can now be identified at the regional level (at the very least, they can be assigned to a state or country). These networks enable visualization and estimation of the correlation between networks and regional innovation dynamics.

Figure 11 encapsulates the results of a first exploration of regional networks in the U.S. from 1975 to 2000. It illustrates the networks of Silicon Valley and Boston in the late 1980s and

graphically demonstrates the argument that California is more highly networked. It (and other work, see Fleming and Marx, 2006) illustrates the connecting importance of educational institutions (such as Stanford and MIT) and corporate postdoctoral fellowships (see Fleming, King and Juda, 2007). Regressions of all 337 U.S. MSA regions demonstrated a correlation between connectedness and subsequent patenting. This work also used the network diagrams to identify crucial cut-point inventors and similar "counter-factual" inventors that did not. Both sets of inventors were then interviewed to provide a structural history of the regions' collaboration dynamics. One possibility for future research would be to replicate these analyses across many U.S. regions. The combination of illustrations, estimations, and field work would greatly facilitate our understanding of regional innovative dynamics. Combined with instrumental variables or natural experiments, the data could illuminate many important policy questions in science and innovation policy. For example, what is the impact of non-compete laws upon mobility, the value of social capital, and knowledge spillovers? Do regions that enforce noncompetes experience a brain and knowledge drain to regions that do not?

A variety of national innovation policy questions could be pursued with the data as well, for example, the influence of international mobility upon knowledge flows and subsequent innovation, or similarly, the influence of international collaboration networks. The impact of foreign investment on the geography of innovation could be studied. Having a long series of geographical data linked to many of the patents will enable a host of longitudinal analyses.

Technology domain level of analysis research that will be enabled by the database:

The data would also afford a better understanding of the emergence of entire fields of technology. New disciplines and breakthroughs are often thought to occur from the intersection of formerly distinct disciplines and technologies. This can only occur, however, through the collaboration of individuals across boundaries or the crossing of individuals across disciplines. One of the biggest challenges to studying this idea, however, is the difficulty of observing – and predicting the success – of such collaborations. Stated another way, it is easy to look backwards and trace the emergence of DNA arrays to a fusion of semiconductor and biotechnology. Given all potential fusions at the time, however, could we have predicted which one would have

occurred – and been successful? By developing the social networks of the entire patenting record, we can avoid sampling on the dependent variable of success. The proposal would enable both illustration and description and inference on these issues. The proposal would also facilitate the measurement of the diffusion of knowledge across these four levels of analysis (individual, organizational, regional, and technological).

Novel econometric techniques (King and Zeng, 2001; Sorenson and Fleming, 2004; Singh, 2005) enable better exploitation of patent citation data to understand knowledge diffusion. These techniques could be combined with the social network data (since much knowledge diffusion occurs with the mobility of researchers, see Singh, 2005). Together, the flow of knowledge could be identified across individual, organizational, regional, and technological boundaries.

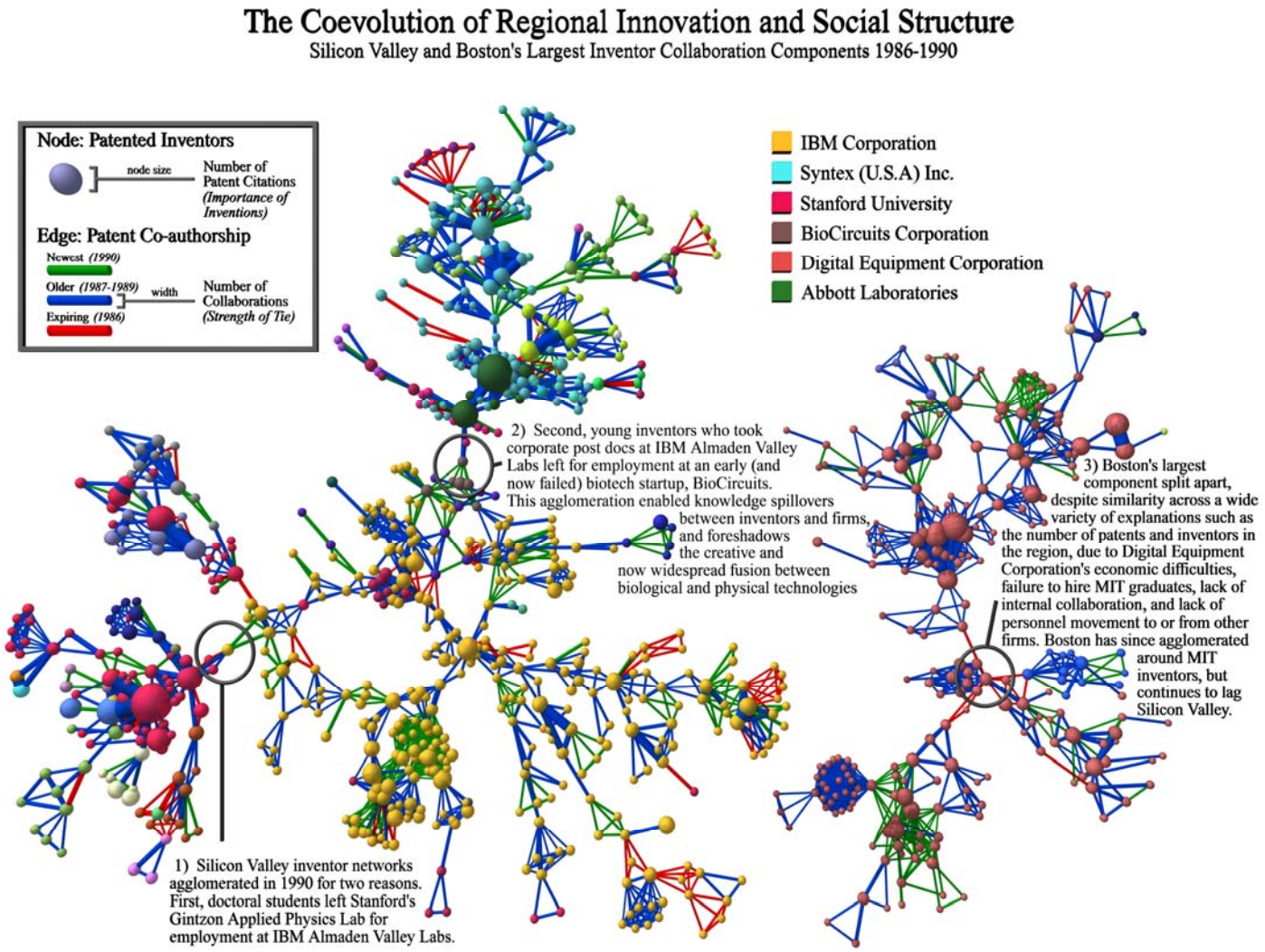


Figure 11: Graphical comparison of Silicon Valley and Boston collaboration networks in 1990 (unpublished, developed with Ivin Baker). The text describes why Silicon Valley networks became connected and Boston did not, based on interviews and field research of the cutpoint inventors (circled) and similar "counter-factual" inventors who did not create cut-points.

Conclusion

Comments and improvements are encouraged.

References

Azoulay, P. and J. Zivin, J. Wang (2007). "Superstar Extinction." Working paper, MIT Sloan School of Management.

Fleming, L. and A. Juda, "A Network of Invention," *Harvard Business Review* 82 (2004): 6.

Fleming, L. and C. King, A. Juda, "Small Worlds and Regional Innovation." *Organization Science*, Vol. 18, No. 2 (2007), pp. 938-954.

Fleming, L. and S. Mingo, D. Chen (2007). "Collaborative Brokerage, Generative Creativity, and Creative Success." *Administration Science Quarterly,* 52 (2007): 443-475.

Fleming, L., and D. Waguespack (2007). "Boundary spanning, brokerage, and the emergence of leadership in open innovation communities." Organization Science, Vol. 18: 165-180.

Granovetter, M. (1973) "The strength of weak ties." American Journal of Sociology, 78: 1360-1379.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. (2001). The NBER patent Citations Data File: Lessons Insights and Methodological Tools, NBER.

Herzog, T. and F. Scheuren, W. Winkler (2007). Data Quality and Record Linkage Techniques. New York, Springer Press.

Marx, M., Strumsky, D., and L. Fleming (forthcoming). "Mobility, Skills, and the Michigan Non-compete Experiment," to appear in *Management Science*.

Marshall, A. *Industry and Trade.* London, MacMillan, 1919.

Owen-Smith, Jason & Walter W. Powell (2004) "Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community." *Organization Science.* 15(1):5-21

Saxenian, A. *Regional Advantage*. Cambridge, MA: Harvard University Press, 1994.

Singh, J. 2005 "Collaborative networks as determinants of knowledge diffusion patterns." Management Science, 51(5): 756–770.

Singh, J. (2007). "Asymmetry of Knowledge Spillovers between MNCs and Host Country Firms." *Journal of International Business Studies*, 38(5): 764-786.

Singh, J. (2008). "Distributed R&D, Cross-regional Knowledge Integration and Quality of Innovative Output." *Research Policy*. *Research Policy,* 37(1): 77-96.

Sorenson, O., and L. Fleming 2004 “Science and the diffusion of knowledge.” Research Policy, 33(10): 1615-1634.

Torvik, V. and M. Weeber, D. Swanson, N. Smalheiser (2005). “A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation,” JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 56(2):140–158, 2005.

Trajtenberg, M., G. Shiff, and R. Melamed (2006). The Names Game: Harnessing Inventors Patent Data for Economic Research, NBER.

Appendix 1
Example of a single USPTO XML file

Introduction

XML stands for Extensible Markup Language. It is a format that has gained popularity due to its minimal size, structure, flexibility and ease of understanding. Data items within an XML file are defined through the usage of tags and attributes.

XML example through HTML based websites

For those who have created basic HTML based websites, the terms of tag and attributes may be easy to understand. For those who have not, I have created a basic example below to demonstrate the basic structure. This initial note is not meant to serve as a XML primer. For a comprehensive overview, there are several online resources to enhance understanding.

```
<html>
    <head>
        <title>This is a Title</title>
    </head>
    <body>
        <img src="picture.jpg" width="100px" />
    </body>
</html>
```

Tags indicate the start and end of information. In our example, we find several tags. Examples of start tags are `<html>`, `<head>`, `<title>`, and `<body>`. End tags are identified through usage of a backslash "`/`" and appear in our example as `</html>`, `</head>`, `</title>` and `</body>`. If we concentrate on the `<title>This is a Title</title>`, the tags indicate that the variable "`title`" contains the value "`This is a Title`". The `<img>` (image) tag contains both start and end while containing separate attributes, namely `src` and `width`. These attributes define the image.

Tags can further separate data into parent-child like structures and are evident through our example. The `<title>` tag is enclosed within the `<head>` tag which creates a relationship between the two variables. This allows us to determine that `<title>` is the child of `<head>` and `<head>` is a parent to `<title>`.

If this example were a HTML page it would communicate to a modern browser that the title of the page should be labeled as "`This is a Title`" and the body of the browser should contain an

image with a width of 100px and the filename picture.jpg. This powerful but simple structure allows both information and relationships to be easily determined.

USPTO XML file formats and data

XML files are evident within the USPTO weekly XML files. A word of caution- USPTO creates a XML file for each unique patent. Since every week, there are several hundred patents being granted, USPTO consolidates the XML file into one file. The structure is not ideal, but can be easily separated to allow for easier data manipulation.

We have taken a sample XML file from the first patent granted on December 30, 2008 and will describe how it aids in the dataset creation process. To increase the efficiency we have determined the creation of smaller, independent datasets is a reasonable approach.

Several datasets are created through from the USPTO patent data including: assignees (patent owner information), citations (referenced patent that aided in development), classes (patent classification scheme), inventors (demographic information about the authors of the patent) and finally, patent (basic chronological information to describe the patent). There are other forms of information which can be found within the XML file, but these items serve as the basis for our current iteration of the inventor disambiguation algorithm.

```
<publication-reference>
    <document-id>
        ……
        <doc-number>D0583526</doc-number>
        ……
    </document-id>
</publication-reference>
```

The unique identifier for each patent is defined through the document number. The document number in our current patent is D0583526. The patent type has been defined as “D” or “Design” and the patent number is 583526 netting the document number D|0583526. The combination of patent type and number will be associated with each of these individual datasets.

USPTO XML | Assignee

```
<assignees>
    <assignee>
        <addressbook>
            <orgname>Kellogg Company</orgname>
            <role>02</role>
            <address>
```

```
                <city>Battle Creek</city>
                <state>MI</state>
                <country>US</country>
            </address>
        </addressbook>
    </assignee>
</assignees>
```

Assignee information can be easily determined from the XML extract. Specifically, Kellogg Company which is located in Battle Creek, MI USA with a role or type defined as 2. From our understanding, assignee type further classifies the firm, but we have yet to receive clarification.

USPTO XML | Citations

```
<references-cited>
    <citation>
        <patcit num="00001">
            <document-id>
                ……
                <doc-number>D20662</doc-number>
                ……
            </document-id>
        </patcit>
        ……
    </citation>
    ……
    <citation>
        <patcit num="00038">
            <document-id>
                ……
                <doc-number>D540507</doc-number>
                ……
            </document-id>
        </patcit>
        ……
    </citation>
</references-cited>
```

The development of patents is oftentimes based on research from previous patents, known simply as a citation. Within the USPTO XML file, citations are organized numerically through a patent citation number reference. The document number mimics the format that constitutes the previously mentioned patent number and type. Here we are able to determine that this patent's first citation is to design patent #20662 and its 38th citation is to design patent #540507.

USPTO XML | Classes

```
<classification-national>
    <country>US</country>
    <main-classification>D 1128</main-classification>
</classification-national>
```

Patent classification can be derived from the main classification field into a class (D1) and subclass (128) resulting in the entry “D 1|128”. Oftentimes, there can be multiple classifications for a patent. If this is the case, we attribute the first classification as primary for the patent.

USPTO XML | Inventors

```
<applicants>
    <applicant sequence="001" app-type="applicant-inventor" ……>
        <addressbook>
            <last-name>Roach</last-name>
            <first-name>Richard</first-name>
            <address>
                <city>Schaumburg</city>
                <state>IL</state>
                <country>US</country>
            </address>
        </addressbook>
        ……
    </applicant>
    ……
    <applicant sequence="005" app-type="applicant-inventor" ……>
        <addressbook>
            <last-name>Barnes</last-name>
            <first-name>Donald</first-name>
            <address>
                <city>Augusta</city>
                <state>MI</state>
                <country>US</country>
            </address>
        </addressbook>
        ……
    </applicant>
</applicants>
```

Several applications or inventors are associated with each patent. The USPTO XML file provides basic information on the inventors including their names, general location, and their sequence within each patent. For our example, we are able to determine that the first author or inventor of the patent is Richard Roach from Schaumburg, IL US. The fifth inventor is Donald Barnes from Augusta, MI US. This information is further consolidated and forms the basis of our inventor disambiguation algorithm.

USPTO XML | Patents

```
<publication-reference>
    <document-id>
        ……
        <date>20081230</date>
    </document-id>
</publication-reference>

<application-reference appl-type="design">
    <document-id>
        ……
        <date>20070730</date>
    </document-id>
</application-reference>
```

Finally, we are interested in considering patent information to understand the timing of the patent. The publication reference provides the week (12/30/2008) when the patent was granted and the application reference provides the week (7/30/2007) when the patent was applied.

A simple search of this patent[21] confirms the information presented within this Appendix. Below, we have provided the design patent as it appears on USPTO's website as of January 7, 2009.

```
United States Patent       D583,526
Roach,   et al.            December 30, 2008

Description:  Curved trapezoid food product
Claim:        We claim the ornamental design for the curved trapezoid food product, as
              shown and described.
Inventors:    Roach; Richard (Schaumburg, IL), Almeida; Helbert (Battle Creek, MI),
              Anderson; Brian (Augusta, MI), Howrey; Bruce (Battle Creek, MI), Barnes;
              Donald (Augusta, MI)
Assignee:     Kellogg Company (Battle Creek, MI)
Appl. No.:    D/282,813
Filed:        July 30, 2007

Current U.S. Class:           D1/128
Current International Class:  0101
Field of Search:              D1/100-130, 199 426/94-95, 89, 128, 144, 293, 543, 549-
                              550, 556, 383, 496, 438-439, 446, 450, 502-504, 512,
                              516, 619, 805, 808

U.S. Patent Documents
   1.   D20662      April 1891          Pearson
   2.   D22990      December 1893       Mackey
   3.   D31777      October 1899        Fox
```

21 Users are able to search for patents: http://patft.uspto.gov/netahtml/PTO/search-bool.html

4.	3384495	May 1968	Potter, Jr.
5.	D212542	October 1968	McCarthy
6.	D213945	April 1969	Cooper
7.	D219002	October 1970	Gronberg
8.	3545979	December 1970	Ghafoori
9.	D247071	January 1978	Neidenberg et al.
10.	D268539	April 1983	Hamann
11.	D273814	May 1984	Gellman et al.
12.	D286919	November 1986	Flockhart
13.	D311472	October 1990	Giles
14.	D324290	March 1992	Stein
15.	D328964	September 1992	Karppinen
16.	D343494	January 1994	Thorniley
17.	5366749	November 1994	Frazee et al.
18.	D353032	December 1994	Mistretta
19.	D376039	December 1996	Pike
20.	D395535	June 1998	Reichkitzer
21.	5843503	December 1998	Clanton
22.	D403485	January 1999	Clanton
23.	D421827	March 2000	Doyle
24.	6120827	September 2000	Rocca
25.	D443953	June 2001	Slaboden
26.	D445237	July 2001	Boselli et al.
27.	D445992	August 2001	Reinhart
28.	D446254	August 2001	Azar
29.	6387421	May 2002	Clanton
30.	D487951	April 2004	Barry et al.
31.	D490590	June 2004	Ferguson et al.
32.	D498034	November 2004	Schwartzberg et al.
33.	D503771	April 2005	Costa
34.	D506302	June 2005	Schwarzberg
35.	D516668	March 2006	Costa
36.	D518272	April 2006	Schwartzberg
37.	D532181	November 2006	Almeida
38.	D540507	April 2007	Aleman et al.

Other References
Golden Stoneground Wheat Crackers packaging. cited by examiner.

Primary Examiner: Webster; Robin V
Assistant Examiner: Kearney; Karen E
Attorney, Agent or Firm: Dickinson Wright PLLC

The XML files contain a tremendous amount of detail and can soon be overwhelming. To better understand the file and terminology, we approach the USPTO and through correspondence we were able to determine clarification. We have provided this clarification in Appendix 2.

Appendix 2
Further USPTO XML file clarification

This Appendix is based upon correspondences with the USPTO and further clarifies the XML patent file. As of July 18, 2008, the USPTO has not created a formal data dictionary for the XML file but we have received some clarification. The formal creation of the data dictionary is an item which the USPTO realizes and may create in the near future.

Only the presentation of the information has been altered to remain consistent with the paper. The language provided is what has been provided by the USPTO and has not been altered.

Table 1 - U.S. Patent Grant and Published Applications Document Numbers:

Patent Grant Patent Number

- Design Patents
    - Position 1 – A constant "D" identifying the granted document as a Design Patent.
    - Positions 2-8 – Seven-position numeric, right justified, with a leading zero.
- SIR Patents
    - Position 1 – A constant "H" identifying the granted document as a Statutory Invention Registration (SIR).
    - Positions 2-8 – Seven-position numeric, right justified, with a leading zero.
- Plant Patents
    - Positions 1-2 – A constant "PP" identifying the granted document as a Plant Patent.
    - Positions 3-8 – Six-position numeric, right justified, with a leading zero.
- Reissue Patents
    - Position 1-2 – A constant "RE" identifying the granted document as a Reissue Patent.
    - Positions 3-8 – Six-position numeric, right justified, with a leading zero.
- Utility Patents
    - Positions 1-8 – Eight-position numeric, right justified, with a leading zero.
- X-Series
    - Patents issued between July 31, 1790 and July 4, 1836. They were not originally numbered, but have since been assigned numbers in the sequence in which they were issued
    - Positions 1-8 – Eight-position, right justified, with a leading "X".

Table 2 - U.S. Patent Grants and Patent Published Applications

Kind Codes

Note: The following 2-position kind codes will be present in the XML <kind> tags of Red Book and Yellow Book. These 2-positions kind codes will also be present on the printed documents with the following exceptions: Reissues will contain a single position "**E**", SIR documents will contain a single position "**H**", and Designs will contain a single position "**S**".

- A1 - Utility Patent Grant issued prior to January 2, 2001.
- A1 - Utility Patent Application published on or after January 2, 2001
- A2 - Second or subsequent publication of a Utility Patent Application
- A9 - Correction published Utility Patent Application
- Bn - Reexamination Certificate issued prior to January 2, 2001.
    - NOTE: "n" represents a value 1 through 9.
- B1 - Utility Patent Grant (no published application) issued on or after January 2, 2001.
- B2 - Utility Patent Grant (with a published application) issued on or after January 2, 2001
- Cn - Reexamination Certificate issued on or after January 2, 2001.
    - NOTE: "n" Represents a value 1 through 9 denoting the publication level.
- E1 - Reissue Patent
- H1 - Statutory Invention Registration (SIR) Patent Documents.
    - Note: SIR documents began with the December 3, 1985 issue
- I1 - "X" Patents issued from July 31, 1790 to July 13, 1836
- I2 - "X" Reissue Patents issued from July 31, 1790 to July 4, 1836
- I3 - Additional Improvements – Patents issued issued between 1838 and 1861.
- I4 - Defensive Publication – Documents issued from Nov 5, 1968 through May 5, 1987
- I5 - Trial Voluntary Protest Program (TVPP) Patent Documents
- NP - Non-Patent Literature
- P1 - Plant Patent Grant issued prior to January 2, 2001
- P1 - Plant Patent Application published on or after January 2, 2001
- P2 - Plant Patent Grant (no published application) issued on or after January 2, 2001
- P3 - Plant Patent Grant (with a published application) issued on or after January 2, 2001
- P4 - Second or subsequent publication of a Plant Patent Application
- P9 - Correction publication of a Plant Patent Application
- S1 - Design Patent

Table 3 - U.S. Application Series Codes

Code:	Filing Dates:
02	Filed prior to January 1, 1948
03	January 1, 1948 through December 31, 1959
04	January 1, 1960 through December 31, 1969
05	January 1, 1970 through December 31, 1978
06	January 1, 1979 through December 31, 1986
07	January 1, 1987 through January 21, 1993
08	January 22, 1993 through January 20, 1998
09	January 21, 1998 through October 23, 2001
10	October 24, 2001 through November 30, 2004
11	December 1, 2004 through December 5, 2007
12	December 6, 2007 through Current

Design Patents

Code:	Filing Dates:
07	Filed prior to October 1, 1992
29	Filed after October 1, 1992

Note: The Design Series Coded "29" is present in the XML data as "29" and is displayed as a "D" on Patent on the Web.

Table 4 - U.S. Patent Classifications

Class

- A 3-position alphanumeric field right justified with leading spaces.
- Design Patents
    - The first position will contain a "D".
    - Positions 2 and 3, right justified, with a leading space when required for a single digit class.
- Plant Patents
    - Positions 1-3 will contain a "PLT"
- All Other Patents
    - Three alphanumeric positions, right justified, with leading spaces

Sub-Class

- Three alphanumeric positions, right justified with leading spaces, and, if present, one to three positions to the right of the decimal point (assumed decimal in the Red Book XML), left justified.
- A digest entry as a sub-class would appear as follows:
    - Three positions containing “DIG”, followed by one to three alphanumeric positions, left justified.

Appendix 3
Dataset Data dictionaries

The primary dataset which contains the disambiguation algorithm is the consolidated inventor dataset. Other supporting datasets contribute either to creating the consolidated inventor dataset or enhance the algorithm.

Table 1: Primary Dataset – Consolidated Inventor Dataset

This dataset is a consolidation of the original inventor dataset and supporting datasets such as assignee, patent, and classes. The inventor disambiguation algorithm processes the data within this dataset and it is the most relevant with regards to this paper. Variables generated from the algorithm include Invnum, Invnum_N, USFlg, Loc, and SeqID.

Variable	Type	Format	Description
Patent	Number		8 digit number identification assigned by the USTPO
Pat_Type	String	1 char	Patent Type (U = Utility, D = Design, H = Statutory Invention Registration, P = Plant, R = Reissue, X = X-Series)
Assignee	String	85 chars	Name of the owner of the patent
AssigneeType	Number		Numerical (1-16) USPTO classification type for owner (values are currently unexplained by USPTO)
AppDate	Date	MMYYDD10.	Week of patent application
AppYear	Number	4 digits	Year of patent application
USFlg	Number	0 or 1	Whether inventor is located in the US
City	String	35 chars	Primary City of inventor
Loc	String	2 chars	If US location, state code. Foreign, country code.
Zipcode	String	5 chars	Zipcode, relevant only for US Locations.
X	Number	11 digits, 6 decimals	Longitude (degrees) of the center (centroid) of ZIP code.
Y	Number	11 digits, 6 decimals	Latitude (degrees) of the center (centroid) of ZIP code.
Invnum	Number		Unique Inventor Identifier for specific record
Invnum_N	Number		Unique Inventor Identifier after match algorithm
Firstname	String	25 chars	Inventor Firstname
Middlename	String	20 chars	Inventor Middlename
Lastname	String	35 chars	Inventor Lastname
Othername	String	20 chars	Inventor Othername
Suffix	String	4 chars	Inventor Suffix such as "JR"

Inv_Seq	Number		Patent Author sequence. (1 = lead author)
Inv_F	String	75 chars	Co-Inventors Firstname (Up to 4, separated by "|")
Inv_L	String	75 chars	Co-Inventors Lastname (Up to 4, separated by "|")
Class	String	75 chars	Primary patent classification (Up to 4, separated by "|"). Secondary classification not included.

Table 2: Supporting Dataset – Assignee Dataset

The assignee dataset has undergone minor updates, mostly to adjust string characters to conform to basic standards. These standards include removing excess whitespace, removal of tags, and translation of Unicode characters. There are additional concepts which have been applied to the primary inventor dataset such as defining a "Loc" variable which can also apply to the assignee dataset. The variable AssigneeID is USPTO's unique identification for assignees and because records have been found to contain blank (occurs, 57% of the time) or inconsistent values, it will not be implemented into the algorithm until it is better understood. Geographic information regarding the assignee has been found to be over 60% of the time missing, so it is not contained within the original dataset. The primary assignee dataset is incorporated (assignee name, assignee type) into the inventor dataset to aid in the disambiguation algorithm. No additional variables have been generated.

Variable	Type	Format	Description
Patent	Number		8 digit number identification assigned by the USTPO
Pat_Type	String	1 char	Patent Type (U = Utility, D = Design, H = Statutory Invention Registration, P = Plant, R = Reissue, X = X-Series)
Assignee	String	85 chars	Name of the owner of the patent
AssigneeID	Number		USPTO's unique identification for Assignees.
AssigneeType	Number		Numerical (1-16) USPTO classification type for owner (values are currently unexplained by USPTO)
Ass_Seq	Number		Patent Assignee sequence (1 = primary assignee)
City	String	35 chars	Primary City of assignee
State	String	2 chars	Primary State code of assignee (US only)
Country	String	5 chars	Primary Country code of assignee

Table 3: Supporting Dataset – Citations Dataset

The citation dataset depends primarily on the Patent, Pat_Type, Citation, and Cit_Type. Other variables such as the Cit_Date, Cit_Name, and Kind are not being applied and are redundant with fields found within the patent dataset. The fields have been preserved to remain consistent with USPTO's structure. The network created from patents and citations is utilized to increase the number of match pairs determined. No additional variables have been generated.

Variable	Type	Format	Description
Patent	Number		8 digit number identification assigned by the USTPO
Pat_Type	String	1 char	Patent Type (U = Utility, D = Design, H = Statutory Invention Registration, P = Plant, R = Reissue, X = X-Series)
Citation	Number		Patent number that is cited by the defined patent.
Cit_Type	String	1 char	Patent type cited by defined patent.
Cit_Date	Date	MMDDYY!0.	Patent grant date cited by defined patent
Cit_Name	String	35 chars	Patent primary inventor surname cited by defined patent
Kind	String	2 chars	Patent kind codes (defined in Appendix 2) cited by defined patent

Table 4: Supporting Dataset – Classes Dataset

The classes dataset is based upon a DVD provided by the USPTO known as CASSIS. This DVD has been purchased because classifications have been known to change throughout time, and this is the most comprehensive source that reflects such changes. The DVD is released on a lagged basis and we have employed the usage of August 2008's DVD in this current analysis. The primary class and the up to three subsequent classes are used to support the disambiguation algorithm. No additional variables have been generated.

Variable	Type	Format	Description
Patent	Number		8 digit number identification assigned by the USTPO
Pat_Type	String	1 char	Patent Type (U = Utility, D = Design, H = Statutory Invention Registration, P = Plant, R = Reissue, X = X-Series)
Class	Number	1 char	General patent classification from the USPTO.
Subclass	Number	MMDDYY!0.	More detailed classification from the USPTO
Primary	Number	0 or 1	A primary value of 1 dictates the senior classification.

Table 5: Supporting Dataset – Patent Dataset

The patent dataset contains basic information such as when the patent was applied for and when the patent was granted. This information allows us to put a timeline on the patents and so application timeline information is incorporated to assist in the sorting algorithms within the adjacency matching algorithm. No additional variables have been generated.

Variable	Type	Format	Description
Patent	Number		8 digit number identification assigned by the USTPO
Pat_Type	String	1 char	Patent Type (U = Utility, D = Design, H = Statutory Invention Registration, P = Plant, R = Reissue, X = X-Series)
AppDate	Date	MMYYDD10.	Week of patent application
AppYear	Number	4 digits	Year of patent application
GDate	Date	MMYYDD10.	Week of patent grant status
GYear	Number	4 digits	Year of patent grant status
Kind	String	2 chars	Patent kind codes (defined in Appendix 2) cited by defined patent
Class	String	3 chars	General patent classification from the USPTO.

}