"Matchmaker, Matchmaker, Make Me a Match"

A General Personal Computer-Based Matching Program for Historical Research

Jeremy Atack
University of Illinois at Urbana-Champaign
and *National Bureau of Economic Research*

Fred Bateman
University of Georgia at Athens
and *Indiana University at Bloomington*

Mary Eschelbach Gregson
University of Illinois at Urbana-Champaign

Cross-sectional studies cannot illuminate the process of economic and social change. Even taking a cross-section of the same population or locale at different times provides only a partial answer. The greatest gains come from pooling cross-sectional and time-series data to generate a panel in which multiple observations for individuals are linked together through time. Social scientists have long recognized the value of contemporary panel data from such sources as the Bureau of Labor Statistics, but only recently have we begun constructing historical panel samples.[1]

Our desire to paint a dynamic picture of rural development in the North and West during the nineteenth century brought us to our current project. We are now about halfway through our project of extending the Bateman-Foust sample from the 1860 manuscript censuses of population and agriculture for northern U.S. townships and linking those data to the 1880 censuses for the same communities. Eventually we will link back to the 1850 and 1870 censuses as well.[2] Our primary goal is to provide a consistent, linked, and computer-readable time-series database for agricultural and rural development covering much of the second half of the nineteenth century. These data promise new insights into the development of the American economy at a critical point in its history.

Linking hundreds of thousands of census records by hand is a daunting task, however. To facilitate our project, we developed a general record-linkage program for use on personal computers. Our initial experiments with PC Matchmaker have been successful. There has been no evidence of a serious tradeoff between reliable linkage and linkage time: automated record linkage on a personal computer is accurate and efficient for historical research. Furthermore, PC Matchmaker is usable for nearly any project requiring one-to-one linkage of records from independently generated sources. After simple—and inexpensive—transformation of raw, encoded data, PC Matchmaker can link census records to other census records, census records to tax records, tax records to voting records, or, indeed, virtually any type of record to any other.

THE ATACK-BATEMAN SAMPLE

Although we designed PC Matchmaker as a generalizable tool, in this article we just illustrate its possibilities with examples from our current project. The Atack-Bateman sample—our extension of the Bateman-Foust sample—will include a complete linked transcription of the manuscript schedules of the 1860 and 1880 censuses of population and agriculture for 140 northern U.S. townships. Figure 1 maps the locations of the counties from which the sample townships were chosen. The 102 cross-hatched counties are the original townships included in the Bateman-Foust sample. The additional counties in Connecticut, Delaware, Maryland, Massachusetts, Michigan, New York, Ohio, Pennsylvania,

FIGURE 1
Locations of Atack-Bateman Sample Townships

and Vermont were originally picked to be part of that sample but were dropped because of the expense. The sample townships in the states farther west are new but were chosen by the same process used to select the original set of sample counties and townships.

Data from the population and agriculture schedules for the 1880 census sample (and the thirty-eight new 1860 sample townships) are being keyed directly into database files from microfilm copies. The 1880 census asked twenty-six questions regarding the person, civil condition, occupation, health, education, and nativity of everyone enumerated. The 1860 census asked fourteen similar questions. The agricultural censuses contain information on farm value, labor, capital, acreage, and the output of crops together with the name of the farm operator. The agriculture and population records are then linked together using PC Matchmaker. Finally, the population records are linked across census years.

The records from the 1860 censuses in the original Bateman-Foust sample were linked by hand but were computer coded without names. Fortunately the Bateman-Foust worksheets were microfilmed, so it has been possible to recover the names of the heads of household and farm operators. This permits us to compare hand linkage circa 1970 with machine linkage today as well as to perform machine linkage between the 1860 and 1880 records.[3]

THE STRUCTURE OF PC MATCHMAKER

Our linkage program draws upon the accumulated experiences of historians, genealogists, medical professionals, and government agencies.[4] The program is unique, however, in that it provides a simple-to-use environment for record linkage within the confines of a personal computer.

PC Matchmaker operates in a DOS environment. The main portions of the program are written in QuickBASIC 4.5. Some utilities and subroutines are coded in Microsoft C 6.0. QuickBASIC was chosen not because it is the most efficient language for record linkage but because it is widely known.[5] Should they be needed, modifications to the program can thus be made inexpensively by inexpert users.

No extraordinary hardware or software is required to link files with PC Matchmaker. Nonetheless, considerable disk space is necessary to link even moderately sized files. To avoid random-access memory constraints, PC Matchmaker stores its tables and temporary files on disk. The size of these temporary tables and files cannot be determined a priori—it depends entirely upon the exact nature of the data to be linked. Appendix A gives examples of the size of files linked and the amount of disk space used, but it is impossible to say that these examples are "typical."

The structure and flow of PC Matchmaker is pictured in figure 2. Our program structure is adapted from the Generalized Iterative Record Linkage System described in Howe and Lindsay (1981) and Hill and Pring-Mill (1985). The following sections briefly describe the input, procedures, and output at each stage.

Data Files, Matching Instruction, and Preprocessing

PC Matchmaker understands any dBASE database file. Data for a linkage project can be collected in dBASE format, as we are doing with the Atack-Bateman sample, but most formats can be converted to dBASE using readily available transfer utilities.

PC Matchmaker can compare any two fields of the same dBASE type (i.e., numeric, text, or logical).[6] These fields need not have the same name. Consider, as an example, linking the 1860 and 1880 population schedules with file structures as in table 1. PC Matchmaker can compare the field BPLACE in the 1860 file with BIRTHPLACE in the 1880 file. It can also be told to compare the 1860 field MI with the 1880 field INITIAL.

The most important stage of the linking process—and one that requires considerable thought before running PC Matchmaker—is deciding on criteria for making the links. PC Matchmaker prompts the researcher to provide a Matching Instruction File containing all required directions. Figure 3 shows such a file for the 1860-to-1880 population census linking.[7] PC Matchmaker needs researcher input for three optional functions—*select, block, filter*—and one required function—*match*.

The *select* function of PC Matchmaker's Preprocessor lets the researcher link subsets of the Data Files. For example, in the 1860-to-1880 Matching Instruction File shown in figure 3, we tell PC Matchmaker to link only heads of households in 1860 with heads of households in 1880.[8] Furthermore, we include only people over eighteen years old in 1880, reasoning that younger people should not have been present in 1860. PC Matchmaker itself imposes no restrictions on this optional select function. It is up to the researcher to determine whom to select if the entire file is not to be linked.

The Preprocessor also generates codes for blocking the data. That is, rather than try to compare every record in Data File A with every record in Data File B, the researcher can have PC Matchmaker partition the data first. PC Matchmaker then looks for links only within each partition (*block*), thus limiting the number of comparisons, the time needed, and the amount of researcher intervention. Blocking, however, is not necessary for PC Matchmaker to link records. In fact, Schofield (1990) argues against it.

Names represent the principal means of record linkage for most historical records. In our case, they represent the only means of linking the population and agri-

FIGURE 2
The Structure and Flow of PC Matchmaker

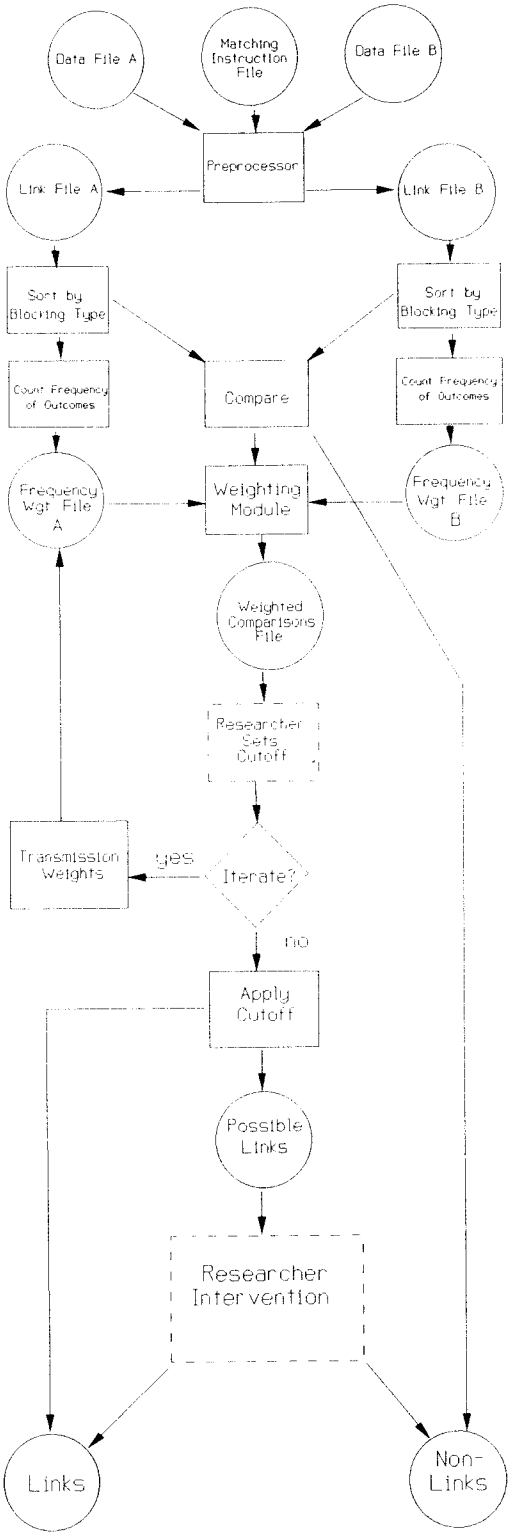

TABLE 1
dBASE File Structures for Atack-Bateman Sample Data Files

1860 data file			1880 data file		
Field name	Type	Width	Field name	Type	Width
FARMNUM	Numeric	5	FARMNUM	Numeric	5
POPPAGE	Numeric	3	PAGE	Numeric	3
POPLINE	Numeric	2	LINE	Numeric	2
HOUSENUM	Numeric	4	DWELLING	Numeric	4
FAMILYNUM	Numeric	4	FAMILY	Numeric	4
LASTNAME	Character	20	LASTNAME	Character	18
FIRSTNAME	Character	20	FIRSTNAME	Character	15
MI	Character	1	INITIAL	Character	1
NOTES	Character	20	COLOR	Numeric	1
COLOR	Numeric	1	SEX	Numeric	1
SEX	Numeric	1	AGE	Numeric	2
AGE	Numeric	2	NEWBORN	Numeric	2
LITERACY	Numeric	1	RELATION	Numeric	1
OCCUPATION	Numeric	2	MARITAL	Numeric	1
BPLACE	Numeric	2	MARRY1880	Numeric	1
PARENT	Numeric	1	OCCUPATION	Numeric	3
REAL_VAL	Numeric	6	UNEMPLOYED	Numeric	2
PERS_VAL	Numeric	5	SICK	Numeric	1
SEQUENCE	Numeric	2	BLIND	Numeric	1
BIRTHYEAR	Numeric	4	DEAF	Numeric	1
			IDIOTIC	Numeric	1
			INSANE	Numeric	1
			MAIMED	Numeric	1
			SCHOOL	Numeric	1
			NOT_READ	Numeric	1
			NOT_WRITE	Numeric	1
			BIRTHPLACE	Numeric	2
			FATHER_BP	Numeric	2
			MOTHER_BP	Numeric	2
			NO_FAMILY	Numeric	2
			BIRTHYEAR	Numeric	4

FIGURE 3
Matching Instruction File for Linking 1860 Population Census to 1880 Population Census

Function	Function type (subtype)	1860 data fields	1880 data fields
select:		RELATIONSHIP = "head"	RELATIONSHIP = "head" AGE > 18
block:	NYSIIS	LASTNAME	LASTNAME
filter:	Name 2 common letters	FIRSTNAME	FIRSTNAME
match:	Name Exact	LASTNAME	LASTNAME
	First 7 letters	LASTNAME	LASTNAME
	First 4 letters	LASTNAME	LASTNAME
	NYSIIS	LASTNAME	LASTNAME
	Name Exact	FIRSTNAME	FIRSTNAME
	First 4 letters	FIRSTNAME	FIRSTNAME
	First letter	FIRSTNAME	FIRSTNAME
	Text	MI	INITIAL
	Numeric	BPLACE	BIRTHPLACE
	Numeric	SEX	SEX
	Numeric Exact	BIRTHYEAR	BIRTHYEAR
	±1	BIRTHYEAR	BIRTHYEAR
	±2	BIRTHYEAR	BIRTHYEAR

cultural schedules. Unfortunately, names are often misspelled or spelled differently from year to year. Further complicating this are antique orthography (ranging from florid to indecipherable) and transcription errors (typos). A phonetic blocking system overcomes some of these limitations. Medical recorders developed the Russell Soundex system, a quasi-alphabetic ordering of last names, to group similar-sounding names. In our terminology, these groups are blocks. Originally devised to deal with Anglo-Saxon names, the Soundex system has been found by the U.S. Immigration and Naturalization Service, for example, to work fairly well with all but Oriental names.[9] Dealing with an almost exclusively French-Canadian population, Bouchard, Roy, and Casgrain (1985) preferred, however, to block their data by using a different grouping of consonant clusters and phonemes better suited to the French language.

The accuracy and completeness of the final results are highly dependent upon the blocking system chosen. This topic is discussed thoroughly in Lynch and Arends (1977). The Soundex system, used by the Census Bureau to index censuses, converts a name to a code of one letter (A through Z) derived from the first letter of the name, and three numbers (0 through 6) derived from the next three consonants of the name. For example, Smith/Smyth/Smythe would all go into block S530. Soundex

TABLE 2
Sample Link Files

1860 link file		1880 link file	
RECNUM	NYSIIS[a]	RECNUM	NYSIIS[a]
1088	ARNSTRANG	1452	ANGL
246	ASBARY	204	ARAN
250	ASBARY	23	ARNSTRANG
772	ASBARY	25	ARNSTRANG
1699	AVARNAN	2420	ARNSTRANG
40	BACAR	636	ASBARY
240	BACAR	645	ASBARY
704	BACAR	822	ASBARY
510	BAS	904	ASBARY
855	BAS	1466	ASLY
1182	BAS	1401	ASTAN
1240	BAS	2025	ASTAN
1313	BAS	1710	AVAN
1548	BAS	1966	AVAN
1704	BAS	616	AVARNAN
		889	BACAR
		720	BAG
		1011	BAL
		1546	BALAD
		701	BALAN
		853	BALAR
		2045	BANTAN
		2270	BANTAN
		2051	BARAT
		1324	BARCLY
		1721	BARN
		1864	BARN
		1867	BARN
		1996	BARN
		2074	BARN
		771	BARY
		229	BARY
		480	BARY
		487	BAS
		545	BAS
		1380	BAS
		1385	BAS
		974	BASAN
		861	BASAN
		1104	BASLY

[a]Compare RECNUMS between tables 2 and 3 to determine the original names corresponding to the NYSIIS codes. For example, BAS = Bush.

has drawbacks for nominal record linkage, however. Misspellings are generally not placed in the same block. Consonant blends that sound like single letters result in similar-sounding names being sorted into different blocks (for example, Shmidt and Shmitz). Such names would never be compared. The New York State Identification and Intelligence System (NYSIIS) improves upon this system (Lynch and Arends 1977).[10]

Lynch and Arends have shown that the NYSIIS rules (appendix B) place the largest number of variations of a name in the same block, thus increasing the probability of making a link because comparisons are made only within, and not between, blocks. At the same time, the number of nonsimilar records in a block is kept to a minimum, meaning that the number of unlikely comparisons is also kept to a minimum. For these reasons we use a modified set of NYSIIS rules to block our census data.[11] PC Matchmaker supports blocking on both Soundex and NYSIIS: the program generates phonetic codes by either set of rules when the respective block types are specified. The program also allows the researcher to supply his or her own field for blocking or to skip this function altogether.

The Preprocessor outputs two Link Files corresponding to the two original Data Files. The Link Files contain the field generated or defined by the block function and the record number for all selected records. The Link Files are then sorted by blocking code to gather together all records in each block. Table 2 shows a part of a sorted Link File. Again, in our 1860-to-1880 link example, the records are grouped by NYSIIS code, and comparisons between NYSIIS blocks are not made. Operations performed by subsequent modules of PC Matchmaker retrieve data from the original data files using the information in the Link Files. This saves considerable disk space.

Counting, Comparing, and Weighting

In the Compare Module, PC Matchmaker examines all possible comparisons between records within blocks. In table 2, record number 1088 in the 1860 Data File is compared with records 23, 25, and 2420 in the 1880 Data File. Likewise, records 246, 250, and 772 of the 1860 Data File are each compared with records 636, 645, 822, and 904 of the 1880 Data File. The Compare Module then uses the *filter* function (see figure 3). Filtering eliminates comparisons that were grouped into the same block but are not otherwise similar. In our example, a comparison between Joe Asbury and Paul Asbury would not be considered further and would be filtered out. Records that fail the filter for all comparisons are output into the Non-Link File and are no longer taken into account.

We have filtered comparisons on the name fields. Filtering on names weeds out comparisons when the fields do not have a researcher-defined number of common letters. The common letters must appear sequentially, but they need not be contiguous. In figure 3 we filtered on the FIRSTNAME field by requiring that there be at least two letters in common. A comparison between Joseph and Paul would thus be thrown out, but one between Joseph and John would be kept, as would one between Paul and Allen. When the number of letters in the field is less than the number defined for the name filter type, all comparisons are retained. In our example, any record having only

FIGURE 4
Operation of Match Function for Linking 1860 Population Census to 1880 Population Census

1. Is LASTNAME in 1860 Data File exactly the same as LASTNAME in 1880 Data File? If yes, skip to line 5.
2. Do LASTNAME in 1860 and LASTNAME in 1880 have the first 7 letters the same? If yes, skip to line 5.
3. Do LASTNAMEs have the first four letters the same? If yes, skip to line 5.
4. Does not meet any criterion except blocked the same: Only NYSIIS codes coincide. Record this at line 5.
5. RECORD ANSWER for this *match* type.
6. Is FIRSTNAME in 1860 Data File exactly the same as FIRSTNAME in 1880 Data File? If yes, skip to line 9.
7. Are first four letters of FIRSTNAME the same? If yes, skip to line 9.
8. Is first letter of FIRSTNAME the same? If yes, skip to line 9.
9. RECORD ANSWER for this *match* type.
10. Are MI and INITIAL exactly the same?
11. RECORD ANSWER for this *match* type.
12. Are BPLACE and BIRTHPLACE exactly the same?
13. RECORD ANSWER for this *match* type.
14. Is SEX equal to SEX?
15. RECORD ANSWER for this *match* type.
16. Is BIRTHYEAR equal to BIRTHYEAR? If yes, skip to line 19.
17. Are BIRTHYEARs within 1 year? If yes, skip to line 19.
18. Are BIRTHYEARs within 2 years?
19. RECORD ANSWER for this *match* type.

TABLE 3
Sample Weighted Comparisons File

RECNUM	LAST	FIRST	MI	BIRTH-PLACE	SEX	BIRTH-YEAR	TOTAL WEIGHT
1240	BUSH	TANDY	Q	12	1	1828	
0487	BUSH	T	Q	12	1	1828	
	6.942514	4.601251	7.94	1.59	0.09	6.13	27.309
1088	ARMSTRONG	JAMES	L	12	1	1814	
2420	ARMSTRONG	JAMES	L	12	1	1815	
	7.357552	4.187627	5.24	1.59	0.09	5.31	23.793
0510	BUSH	COLBY		12	1	1811	
1380	BUSH	COLBY		12	1	1812	
	6.942514	7.942514	0.00	1.59	0.09	6.81	23.392
0246	ASBERRY	CALVIN		12	1	1800	
0645	ASBURY	CALVIN		12	1	1800	
	0.000000	8.942514	0.00	1.59	0.09	8.94	19.572
0240	BAKER	RICHMOND		20	1	1845	
0889	BAKER	RIGGS		20	1	1844	
	8.942514	4.594467	0.00	0.95	0.09	4.52	19.112
0855	BUSH	CHRISTOPHER	C	12	1	1824	
1380	BUSH	COLBY		12	1	1812	
	6.942514	4.472313	0.00	1.59	0.09	−13.37	−0.273
0855	BUSH	CHRISTOPHER	C	12	1	1824	
1385	BUSH	P		12	1	1824	
	6.942514	−16.698237	0.00	1.59	0.09	6.62	−1.447
0772	ASBERRY	JAMES	C	20	1	1838	
0645	ASBURY	JAMES	S	20	1	1839	
	0.000000	4.187627	−11.73	0.95	0.09	4.41	−2.078
1182	BUSH	PAULINE	E	20	2	1856	
1385	BUSH	P		12	1	1824	
	6.942514	5.727984	0.00	−2.22	−1.11	−11.46	−2.139
1240	BUSH	TANDY	Q	12	1	1828	
1385	BUSH	P		12	1	1824	
	6.942514	−16.113275	0.0	1.5	0.09	0.00	−7.483
1088	ARMSTRONG	JAMES	L	12	1	1814	
0025	ARMSTRONG	J	W	20	1	1851	
	7.357552	2.144312	−9.66	−2.76	0.09	−11.79	−14.618
1313	BUSH	WILLIAM		12	1	1812	
0487	BUSH	T	Q	12	1	1828	
	6.942514	−10.513362	0.0	1.59	0.09	−13.89	−15.774
1313	BUSH	WILLIAM		12	1	1812	
1385	BUSH	P		12	1	1824	
	6.942514	−11.098324	0.0	1.59	0.09	−14.37	−16.844

an initial in the 1860 FIRSTNAME field would be compared with all records in the corresponding block of the 1880 Data File. Although our example uses only names, it is also possible to filter comparisons based upon a numeric field or to use no filter at all.

Next the program follows the researcher-defined matching instructions, the fourth and last function in the Matching Instruction File. The *match* function asks a series of questions that can be answered about each comparison that passes the filter. For example, when the match function is defined on a name field, the program can search for an exact match, or a match on the first *n* letters, or a match with the blocking field. Figure 4 illustrates how PC Matchmaker follows the match function from the Matching Instruction File in figure 3. The questions are nested in the sense that when a question is answered in the affirmative, the program records the answer and skips to the next field on which the match function is defined. If all questions about a field are answered in the negative, the program records this before proceeding to the next field.

To apply the statistical formulae upon which the linkage decision is made, the program counts the number of times unique outcomes occur in each field on which a match type is defined. This is done for each level of questioning defined in the Matching Instruction File.

Thus, in the 1860-to-1880 census-matching example (lines 1 to 5 of figure 4), PC Matchmaker counts how many times each unique sequence of letters appears in the LASTNAME fields of the two data files. Then it counts how many times each unique group of first seven letters occurs, how many times each unique group of first four letters occurs, and how many times each block value occurs. Consider, for example, the LASTNAME ARMSTRONG. ARMSTRONG is an "outcome" for exact match on the name field. The number of times ARMSTRONG occurs in the 1860 and 1880 Data Files is recorded in a Frequency Weight File. Likewise, ARMSTRO is a match on the first seven letters of the name. The number of times ARMSTRO occurs is also recorded in a Frequency Weight File.

In the Weighting Module, the result of each question asked by the match function is combined with the statistics from the Frequency Weight Files to give a statistical weight to the likelihood that the two records being compared are for the same person. Specifically, Fellegi and Sunter (1969) show that the likelihood that two records contain information about the same person can be expressed as an odds ratio. For this application it is convenient to use binit weights, where

$$\text{binit weight} = \log_2 \frac{\text{probability of true link}}{\text{probability of random comparison}}.$$

Log to the base two ($\log_2$) is convenient because the elements of the weight, one for each field on which the match function is defined, become additive. Therefore, the odds ratio (the statistical weight given to each comparison) can be computed in steps. Each element of the weight is computed from the $\log_2$ of {the number of times an outcome occurs} divided by {the total number of records}, where again an "outcome" is just a specific value from the Frequency Weight Files (see Howe and Lindsay 1981, 102). This is equivalent to the ($\log_2$ of the) conditional probability of each outcome.

Consider the first comparison in table 3. Records for TANDY Q. BUSH and T. Q. BUSH would seem to be a likely match, especially because both records have 1828 as Mr. Bush's birthyear. Statistically the comparison is also highly likely to be a true match. Suppose the probability that a random comparison is a true match in the files from which the sample Weighted Comparisons File in table 3 was drawn is 1/1000 ($=2^{-10}$). The total weight for a comparison (about 27 for Mr. Bush) closely approximates the $\log_2$ of the odds in favor of the comparison's being a true link given the answers to the match function's questions. Therefore PC Matchmaker's operation determined that the odds that records 1240 and 0487 are for the same person are $2^{27}/2^{10} = 2^{17}$, or 125,000:1.

The weights on the individual fields in the records can be interpreted similarly. The probability of randomly finding two records in the Data Files with identical LASTNAMES was about 2^{-3}. The figure is the product of the probability of finding BUSH in File A and in File B. Given that the LASTNAME fields in records 1240 and 0487 exactly equaled BUSH (which in table 3 has a weight of almost 7), the odds that the comparison is a true match improve to nearly 2^4:1 ($=2^7/2^3$), or 16:1. The more rare a value for a field (e.g., having Q as a middle initial), the more the agreement on that field contributes to the total weight. The more common the characteristic (e.g., having BIRTHPLACE = 12 [for Tennessee] or SEX = 1 [for male]), the less the agreement improves the odds that any comparison is a true match. The total weight for each

TABLE 4
Automatic Linking Compared with Hand Linking of the Bateman-Foust Sample

No. of farms	Linked same		Linked differently		Indeterminate		Not compared	
	No.	(% of all farms)	No.	(% of all farms)	No.	(% of (all farms)	No.	(% of (all farms)
17	8	47.06	1	5.88	2	11.76	6	35.29
55	47	85.45	1	1.82	2	3.64	5	9.09
22	8	36.36	1	4.55	5	22.73	8	36.36
80	70	87.50	1	1.25	1	1.25	8	10.00
40	35	87.50	0	0.00	0	0.00	5	12.50
34	28	82.35	1	2.94	1	2.94	4	11.76
287	225	78.40	1	0.35	47	16.38	14	4.88
60	57	95.00	0	0.00	0	0.00	3	5.00
168	151	89.88	0	0.00	10	5.95	7	4.17
193	153	79.27	1	0.52	8	4.15	31	16.06
113	110	97.35	0	0.00	0	0.00	3	2.65
79	59	74.68	2	2.53	6	7.59	12	15.19

comparison, which represents the likelihood that the comparison is a true link, is the sum of the weights computed for each field.

The output of the Weighting Module is the Weighted Comparisons File. This reports all the comparisons that passed the filter. For each comparison, PC Matchmaker gives the contents of the fields used for the match function, the weights computed for each field, and the total weight for the comparison. A sample Weighted Comparisons File is shown in table 3. These comparisons correspond to records from the Sample Link Files shown in table 2.

Cutoff Values and Iteration

At this point the researcher must intervene. As described by Bouchard (1991), in every matching system there is a tradeoff between completeness and accuracy. For some projects, it is necessary that all links be true links. For other projects, it is more important that the greatest number of records are linked automatically. We decided to leave this choice to the researcher. After studying the reports generated by the Weighting Module, the researcher answers PC Matchmaker's prompt for a cutoff value based on his or her judgment of the project and the data.[12]

After examining the Weighted Comparison File and choosing the cutoff value, the researcher must make another decision: whether or not to iterate through the Weighting Module.[13] What would change? Recall that the components of the total weight for a comparison—one for each field on which the match function was defined—are computed from the Data Files. However, there are many possible sources of error in the data.[14] To the extent that the Data Files do not contain perfectly correct information on the population, the weights overestimate the probability that a comparison is a true link. In effect, PC Matchmaker has assumed that the Data File contains only true information. If a comparison is above the cutoff value (the researcher has determined that it is very likely a true match) but doesn't match exactly on every field, most likely there are errors in the data. The weights can be adjusted downward to account for the probability of such errors. PC Matchmaker computes the proportion of comparisons above the cutoff value that have discrepancies. This is the estimated error rate. During iteration the weights are adjusted by (1 − error rate). Iteration can be continued until the adjustment to the weights is arbitrarily small.

When iteration is no longer desired, PC Matchmaker uses the final cutoff value to output two additional files: Definite Links and Possible Links. Because PC Matchmaker is designed for one-to-one linkage, it accepts as a Definite Link only that comparison that had the highest total weight for each record in Data File A. The Possible Links are all remaining comparisons. The researcher then intervenes to accept these comparisons as links or reject them and add them to the file of Non-Links, where they join those records excluded by the filter function.

PERFORMANCE OF PC MATCHMAKER

Of the many questions that could be asked regarding the performance of PC Matchmaker, we will discuss the two that have been our main concerns in linking the Atack-Bateman sample: (1) How accurate are the automatic links? (2) How much time does automation save?

PC Matchmaker versus Hand Matching: Accuracy

When we link records by hand, we often treat the problems we encounter as judgment calls, but as Roger Schofield (1990) has noted:

> If the judgements we make about specific links have any claim to intellectual respectability, we ought to be able to specify the principles on which they are based. If we can do that, we can express those principles in the form of a computer program and get the machine to implement them far more consistently than we can ourselves.

As we developed PC Matchmaker, we hoped the procedures outlined above were a good systematization of our own mental processes. It was encouraging to see that computer matching of the Bateman-Foust sample of the 1860 censuses was just as good—and occasionally better than—the original hand matching.

Table 4 shows the results of computer linking the coded population and agriculture schedules for six Missouri and six Illinois townships from the Bateman-Foust sample. PC Matchmaker made more than 75 percent of the same links that were made by hand without iteration or manual intervention. Our automated linkage rate is in line with the Census Bureau's linkage rate of 70 percent (Belin 1990, 167). The two townships that PC Matchmaker did not handle well were both very small—seventeen and twenty-two farms each. These same townships had the largest percentage of records that did not get compared at all. Again, no comparisons are made when the names of farmers and heads-of-households are not blocked together or when first names of farmers and heads-of-households in the relevant block are not remotely similar.

For large Link Files, a linkage rate of 75 percent or better before considering indeterminate comparisons reduces the problem of linkage to a manageable size. For example, the largest township in the test group had 287 farms; 226 were linked without manual assistance. The relatively large number of indeterminate comparisons in that township arose because nearly 30 percent of the farm operators had the surnames Roberts or Hulen. These indeterminate comparisons can often be resolved quickly because PC Matchmaker has already grouped

TABLE 5
Examples of Runtime for PC Matchmaker

File A		File B				20Mhz 386			4.77Mhz IBM-PC		
No. of records	No. selected	No. of records	No. selected	No. of match types	No. of match fields	Count frequency	Filter[a]	Weight	Count frequency	Filter	Weight
1860											
323	250	1729	314	10	3	0:27:11	0:29:03	0:47:16	2:45:41	3:02:56	6:10:55
102	60	610	102	10	3	0:06:18	0:06:31	0:07:16	0:36:33	0:37:04	0:37:13
203	110	1092	203	10	3	0:15:19	0:15:52	0:23:39	1:23:51	1:24:34	1:24:43
247	193	1552	238	10	3	0:18:25	0:19:04	0:24:46	2:05:45	2:06:41	2:06:50
150	113	874	150	10	3	0:10:05	0:10:29	0:12:33	1:00:45	1:01:24	1:01:33
155	79	862	147	10	3	0:04:32	0:04:37	0:04:51	1:05:07	1:05:47	1:05:50
73	17	337	73	10	3	0:09:06	0:09:21	0:10:57	0:13:35	0:13:49	0:15:23
93	55	520	90	10	3	0:06:09	0:06:21	0:07:08	0:22:43	0:23:26	0:29:41
56	22	297	55	10	3	0:03:45	0:03:51	0:04:01	0:13:21	0:13:37	0:15:00
122	79	664	119	10	3	0:08:52	0:09:07	0:10:26	0:30:10	0:31:01	0:41:17
65	29	330	65	10	3	0:04:36	0:04:44	0:05:04	0:14:07	0:14:29	0:17:01
172	146	355	34	10	3	0:07:40	0:07:50	0:08:35	0:13:49	0:14:26	0:20:54
1880											
365	365	2428	841	10	3	0:45:13	0:48:10	1:25:15	6:30:26	7:16:32	2:15:36
155	155	1004	268	10	3	0:10:35	0:10:40	0:18:57	1:31:23	1:39:43	2:59:52
216	216	1038	323	10	3	0:27:16	0:26:34	0:38:33	2:26:13	2:33:47	3:35:07
437	437	2774	859	10	3	0:57:20	1:04:55	1:40:18	8:14:22	8:43:53	15:32:11
288	288	1682	551	10	3	0:30:24	0:32:43	1:00:07	3:46:44	4:02:40	8:06:05
213	213	1239	385	10	3	0:28:18	0:30:01	0:56:52	2:25:04	2:34:40	4:07:29
1860—>1880											
1729	314	2428	445	13	6	2:04:55	2:13:33	7:29:19	11:44:41	13:20:35	
610	102	1004	184	13	6	0:36:11	0:37:03	0:59:06			
1092	203	1038	204	13	6	1:04:36	1:05:45	1:20:54	(we didn't go on with this test)		
1552	236	2774	535	13	6	1:47:45	1:51:15	3:33:28			
874	150	1682	319	13	6	0:51:33	0:53:57	1:33:16			

[a]We used one filter, as shown in figure 3.

the most likely links together for inspection. In this case, we were left with just 5 percent of the original file (14 of 287 farms) that had to be linked manually, and some of these may not have farmers in the population file.

Most important to us, though, were differences between PC Matchmaker's links and those made by hand. Most of the "linked differently" cases eliminated a farmer without a farm or a farm without a farmer. That is, linking with PC Matchmaker resulted in more links than did manual linkage.

PC Matchmaker versus Hand Matching: Speed

The Bateman-Foust sample required that we link twenty-one thousand families to their farms. Our extension of that sample to 1880 requires that we link at least double that number of families and identify the persisters in the sample communities. The total number of comparisons possible for Bateman-Foust was 17.5 million. Our extension has at least 300 million possible comparisons. If we linked records at the same rate as we did twenty years ago, our current project would be in progress for decades! Manual matching of this sample is impractical.

Linking with PC Matchmaker produces results quickly. As table 5 shows, the 1860 population-to-agriculture automated links were generated in less than an hour for all but the largest files, using a 20 Mhz 386 computer. The three most time-consuming operations in PC Matchmaker are counting the frequency of outcomes, filtering comparisons, and weighting the remaining comparisons. When we linked the 1860 agriculture-and-population schedules on a 20 Mhz 80386 IBM compatible with a 80387 math coprocessor, the largest township —with 78,500 possible comparisons—was processed in under 2 hours. PC Matchmaker processed a township with 4,950 possible comparisons in 20 minutes. As was the case with disk space, there is no easy way to estimate the time needed to link two files—counting, filtering, and weighting are completely data dependent. Nevertheless, runtimes can be sharply reduced through the use of RAM disks or disk-catching systems to expedite disk

reads and writes. Neither of these techniques was used for the results reported here.

Table 5 also shows PC Matchmaker's runtimes for the 1860 agriculture-to-population linkage on an old 4.77 Mhz IBM-PC. While the program ran three to six times slower, the time is still not unreasonable because linkage is done only once for most projects.

The 1880 test townships were considerably larger, ranging from 41,540 possible comparisons to 375,383 possible comparisons. Linking the 1880 agriculture files to the 1860 population files took from 40 minutes to just under 4 hours on the 386 system. On the IBM-PC, the process took from 6 to 33½ hours. We consider the upper limit to be beyond the maximum practical!

Linkage of the 1860 to 1880 population files—the example used earlier—was slower because it used three additional matching fields. In particular, the weighting module was much slower because of the additional number of computations. Nonetheless, five files were linked in less than 30 hours on the 386 machine.

CONCLUDING REMARKS

PC Matchmaker has removed time spent linking—the major impediment in our efforts to broaden the scope of our research. We believe it will prove useful for other researchers. Microlevel information exists for many facets of life in the past, but seldom are all the desirable data recorded in one place. The difficulty and expense of linking data on individuals has sometimes led scholars to use aggregated data to address microlevel issues. For many projects, this is no longer necessary.

Personal computer-based record linkage works. Although for large files record-linkage time can be substantial, PC Matchmaker effectively reduces the resources needed for linkage. We believe it represents a substantial advance in our ability to link census records and has wide applicability to other record-linkage problems. Beta test versions are available from the authors upon request. Final versions will be released into the public domain as soon as possible.

APPENDIX A
Examples of Disk Space Used for PC Matchmaker's Temporary Files and Reports

File A				File B						Kilobytes of disk space used	
Size (Kb)	No. of records	No. selected	% selected	Size (Kb)	No. of records	No. selected	% selected	No. of match types	No. of match fields	Temporary files	Reports/ Output
134.2	323	250	77.40	185.7	1729	314	18.16	10	3	166.8	169.0
44.4	102	60	58.82	65.9	610	102	16.72	10	3	41.9	21.2
86.2	203	110	54.19	117.5	1092	203	18.59	10	3	103.1	88.0
104.4	247	193	78.14	166.7	1552	238	15.34	10	3	100.8	59.1
64.3	150	113	75.33	94.2	874	150	17.16	10	3	66.4	41.7
66.3	155	79	50.97	92.2	862	147	17.05	10	3	56.5	24.5
34.4	73	17	23.29	38.8	337	73	21.66	10	3	23.8	3.4
43.2	93	55	59.14	59.5	520	90	17.31	10	3	42.7	19.9
26.9	56	22	39.29	34.3	297	55	18.52	10	3	23.5	4.3
56.1	122	79	64.75	75.7	664	119	17.92	10	3	54.9	24.6
30.1	65	29	44.62	38.0	330	65	19.70	10	3	31.8	10.0
78.2	172	146	84.88	40.8	355	34	9.58	10	3	40.6	13.1

APPENDIX B
Rules for NYSIIS Phonetic Blocking Scheme

1. If the first letters of the name are
   MAC, change these letters to letters MCC
   KN, change these letters to letters NN
   K, change this letter to letter C
   PH, change these letters to letters FF
   PF, change these letters to letters FF
   SCH, change these letters to letters SSS
   WR, change these letters to letters RR
   RH, change these letters to letters RR
   DG, change these letters to letters GG
   A, E, I, O, or U, change these letters to letter A
2. Drop terminal letter S or Z from all names before coding begins.
3. If the last letters of the names are
   EE, change these letters to letter Y
   IE, change these letters to letter Y
   YE, change these letters to letter Y
   DT, change these letters to letter T
   RT or RD, change these letters to letter D
   NT or ND, change these letters to letter N
   IX, change these letters to letters ICK
   EX, change these letters to letters ECK
4. The first character-code of the NYSIIS code is the first letter of the name after executing rule (1).

5. In terms of a program loop, the "pointer" is now set to the second letter of the name. The following rules are performed for each subsequent letter in the name.
6. Only one of these rules can apply to each letter in the name.
   (a) If blank, go to rule 8.
   (b) If the end-of-string marker, go to rule 9.
   (c) If the current letter is a vowel and equal to EV, change these letters to character-codes AF; otherwise, change the letter to character-code A.
   (d) If the current letter is a Y and it is not the last letter of the name, change this letter to character-code A.
   (e) If the current letter is Q, change this letter to character-code G.
   (f) If the current letter is Z, change this letter to character-code S.
   (g) If the current letter is M, change this letter to character-code N.
   (h) If the current letter is K and if the next letter is N, replace the current letter by character-code N or else replace the current letter by character-code C.
   (i) If the current letter is S and the next letters are CH, change the current letter to character-codes SSA if end of word, or change to character-codes SS if not end of word.
   (j) If the current letter is S and the next letter is H, change SH to character-codes SA if end of word, or change to character-codes SS if not end of word.
   (k) If the current letter is P and the next letter is H, change PH to character-codes FF.
   (l) If the current letter is G and the next two letters are HT, change GHT to character-codes TTT.
   (m) If the current letter is D and the next letter is G, change DG to character-codes GG.
   (n) If the current letter is W and the next letter is R, change WR to character-codes RR.
   (o) If the current letter is H and either the preceding or following letter is not a vowel, replace the current letter with the preceding character-code.
   (p) If the current letter is W and the preceding character-code is a vowel, replace the current letter with the preceding character-code.
   (q) If none of these rules apply, retain the current letter as the character-code.
7. If the current character-code is equal to the previous character-code, remove the current character code (i.e., no doubles).
8. Move the pointer to the next letter and return to rule 6.
9. If the terminal character-code is S, remove it.
10. If the terminal two character-codes are AY, replace them with the character-code Y.
11. If the terminal character-code is A, remove it.

Source: Adapted from Lynch and Arends (1977), Appendix B.

NOTES

1. Analysis of economic and demographic issues using linked census data by Barron (1984), Burton (1985), Curti (1959), Faragher (1986), and others has shown that the study of a population over time is a fruitful endeavor. Also under way are "total history projects," including the Cambridge family reconstitution project (Schofield 1990), the Edgefield, South Carolina, database (Burton 1985), Philadelphia Social History Project (Condran and Seaman 1981), and the Texas Historical Demography Project (Vetter, Gonzalez, and Gutmann 1990). These inquiries add depth to the study of limited geographical regions. On a wider scale, the Inter-University Center for Population Research has linked all vital records for Saguenay for the period 1842–1971 and is currently linking marriage records for the entire Province of Quebec (Bouchard 1991; Bouchard and De Braekeleer 1991; Bouchard, Roy, and Casgrain 1985). Until now, however, no historian has expanded the scope of study to cover a large and diverse section of the United States.
2. Our current project is supported by NSF SES-8914861 at the University of Illinois at Urbana-Champaign, and SES-9001066 at Indiana University and the University of Georgia at Athens. For a complete description of the original Bateman-Foust sample, see Bateman and Foust (1974).
3. A number of changes took place in township boundaries between 1860 and 1880. These changes complicate the process of data collection, but the problem is not insuperable. We are making every effort to collect the same geographical areas in 1880 as were sampled in 1860. For example, part of Luzerne County, Pennsylvania, was split off to become Lackawanna County in 1878; and our sample township, Abington, was in the portion of the county transferred to Lackawanna (see *The Compendium of the Tenth Census*, page 49). Indeed, the township itself was also divided into North Abington and South Abington. We are collecting both. Similarly, in 1860 (presumably just after the 1860 census was taken), Princeton township in Benton County, Minnesota, was transferred to Mille Lacs County and split into Princeton, Greenbush, and Milo townships (see *Eighth Census, Population of the United States in 1860*, Table 3, pages 177 and 179). We are collecting all three for 1880.
4. The most important sources include Acheson (1967 and 1968), Fellegi and Sunter (1969), Winchester (1970), Wrigley (1973), Condran and Seaman (1981), Howe and Lindsay (1981), Hill and Pring-Mill (1985)—most of which are reprinted in Kilss and Alvery (1985). Kilss and Jamerson (1990) report recent improvements in the record-linkage technique implemented by the Census Bureau.
5. Many thanks to programmer/consultant Timothy A. Gregson for his effort and patience, particularly in accommodating our taste for QuickBASIC.
6. Two caveats: (1) Memo fields are not explicitly supported and should be stripped from data files prior to using the PC Matchmaker, and (2) if dBASE indexing is desired, data files must be reindexed after processing with PC Matchmaker.
7. The user can formulate a Matching Instruction File prior to calling PC Matchmaker or can create one using a menu system that is called automatically when a Matching Instruction File is not specified on the command line.
8. The Atack-Bateman sample will identify as many links as possible for the entire population, not just household heads. Linking on household heads alone, however, gives us additional information to use in linking dependents (but note that the procedure has the same statistical consequences as pretesting).
9. We have several hundred Chinese in our 1880 California samples. It remains to be seen how well our blocking scheme copes with them.
10. We have been unable to trace the exact origins of NYSIIS. All citations go back to Lynch and Arends (1977), but their paper gives no primary source reference. A private conversation with William Arends (USDA) reveals that an unnamed source provided the coding scheme. It is believed that this blocking scheme was developed by New York State to track radicals during the late 1960s and early 1970s. Other intelligence agencies (such as the CIA, NSA, and Treasury) have similar record-linkage programs.
11. The Soundex codes will, however, be generated and included with the distributed final dataset to facilitate other researchers' following of migrants from the townships using the Soundex index provided by the Bureau of the Census. Our version of NYSIIS codes (see appendix B) differs from the modified rules given in Lynch and Arends (1977) by changing DT to D (Lynch and Arends code this as T) at the end of a name, and SCH to SS instead of SSS.
12. The difficulty of automatically calculating the optimal cutoff weight using the Fellegi-and-Sunter method, as we do, is that the assumption of independence between the fields is violated in practice. Winkler (1990) and Belin (1990) propose improvements on the Fellegi-and-Sunter method of calculating optimal cutoff values.
13. Iterative techniques are explained in detail in Howe and Lindsay (1981)—see especially pages 102–3.

14. For example, not only did people rarely spell their names for the census enumerators, sometimes they also supplied faulty information. Consider the phenomenon of "heaping"—the rounding of ages to the nearest five or ten years (Kelly 1974). There is also the problem of transcription error when collecting the data.

REFERENCES

Acheson, E. D. 1967. *Medical record linkage.* Oxford: Oxford University Press.

———, ed. 1968. *Record linkage in medicine.* Edinburgh.

Barron, H. 1984. *Those who stayed behind: Rural society in nineteenth-century New England.* Cambridge: Cambridge University Press.

Bateman, F., and J. D. Foust. 1974. A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History* 48:75–93.

Belin, T. R. 1990. A proposed improvement in computer matching techniques. In *Statistics of income and related administrative record research: 1988–1989*, edited by B. Kilss and B. Jamerson, 167–72.

Bouchard, G. 1991. Current issues and new prospects for computerized record linkage in the Province of Québec. Working paper, Inter-University Center for Population Studies.

Bouchard, G., and M. De Braekeleer, eds. 1991. *Histoire d'un génôme. Population et génétique dans l'est du Québec.* Québec: Press of the University of Québec.

Bouchard, G., R. Roy, and B. Casgrain. 1985. *Reconstitution automatique des familles: Le système SOREP.* Chicoutimi: University of Québec at Chicoutimi.

Burton, V. 1985. *In my father's house.* Chapel Hill: University of North Carolina Press.

Condran, G. A., and J. Seaman. 1981. Linkage of the 1880–81 Philadelphia death register to the 1880 manuscript census: A comparison of hand- and machine-record linkage techniques. *Historical Methods* 14:73–84.

Curti, M. 1959. *The making of an American community: A case study of democracy on a frontier county.* Stanford: Stanford University Press.

Faragher, J. M. 1986. *Sugar Creek: Life on the Illinois prairie.* New Haven: Yale University Press.

Fellegi, I. P., and A. B. Sunter. 1969. A theory for record linkage. *Journal of the American Statistical Association* 64:1183–1210. Reprinted in *Record linkage techniques—1985*, edited by B. Kilss and W. Alvery, 51–78.

Hill, T., and F. Pring-Mill. 1985. Generalized iterative record linkage system. In *Record linkage techniques—1985*, edited by B. Kilss and W. Alvery, 327–36.

Howe, G. R., and J. Lindsay. 1981. A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research* 14:327–40. Reprinted in *Record linkage techniques—1985*, edited by B. Kilss and W. Alvery, 97–110.

Kelly, D. 1974. Linking nineteenth-century manuscript census records: A computer strategy. *Historical Methods Newsletter* 7:72–82.

Kilss, B., and W. Alvery, eds. 1985. *Record linkage techniques—1985.* U.S. Treasury Department Internal Revenue Service, Statistics of Income Division, Washington, D.C.: GPO.

Kilss, B., and B. Jamerson, eds. 1990. *Statistics of income and related administrative record research: 1988–1989.* U.S. Treasury Department Internal Revenue Service, Statistics of Income Division, Washington, D.C.: GPO.

Lynch, B. T., and W. L. Arends. 1977. *Selection of a surname coding scheme for the SRS record linkage system*, Sample Survey Research Branch, Research Division Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

Schofield, R. 1990. Automatic family reconstitution: The Cambridge experience. Paper prepared for SSHA meeting, Minneapolis, October 1990.

Vetter, J. E., J. R. Gonzalez, and M. P. Gutmann. 1990. Computer-assisted record linkage using a relational database system. Paper prepared for SSHA meeting, Minneapolis, October 1990.

Winchester, I. 1970. The linkage of historical records by man and computer: Techniques and problems. *Journal of Interdisciplinary History* 1:107–24.

Winkler, W. E. Frequency-based matching in the Fellegi-Sunter model of record linkage. In *Statistics of income and related administrative record research: 1988–1989*, edited by B. Kilss and B. Jamerson, 161–66.

Wrigley, E. A., ed. 1973. *Identifying people in the past.* London: Edward Arnold.