Improving the Study of Campaign Contributors with Record Linkage

Christophe Giraud-Carrier
Data Mining Lab
Department of Computer Science
Brigham Young University

Jay Goodliffe
Center for the Study of Elections and Democracy
Department of Political Science
Brigham Young University

Bradley Jones
Department of Political Science
University of Wisconsin-Madison

March 1, 2010

ABSTRACT

Up to now, most campaign statistics have been reported at the level of the donation. While these are interesting, one often needs to have information at the level of the donor. Obtaining information at that level is difficult as there is neither an unique repository of donations nor any standard across existing repositories. What political scientists need is an accurate way of grouping, or linking, together donations made by the same donor. In this paper, we describe the technique of record linkage, developed in the field of computer science for such a purpose. We show how it

may be effectively applied in the context of nationwide donation data and report on new, previously unattainable results about campaign contributors in the 2007-2008 election cycle.

1. INTRODUCTION

Record linkage is concerned with recognizing when individuals who are expressed in different records are, in actuality, the same person. In political science, a primary use for this process is in the study of political participation, notably voting and campaign finance. For voting, record linkage may involve linking a voter's earlier vote history in one database to a more recent vote history in another database (e.g., because the voter moved).

There are also times where records must be linked within the same database. Most campaign finance databases, e.g., Federal Election Commission (FEC), report data by donation (or transaction). If the same person has given two different donations to the same candidate in the same election cycle, then there will be two records for that individual (assuming the donations must both be reported). However, it is often interesting to note how much money a person gives overall to a candidate in a given election cycle, or how much an individual gives to all candidates in an election cycle. For example, in the 2008 election cycle, Andrew Howard (3131 Bannock Drive, Provo, Utah) gave $600 to Ron Paul, then $450 to Mike Huckabee, and finally $250 to Mitt Romney.

Without record linkage, any aggregate statistics can only be reported at the level of the donation, e.g., reporting the mean or median donation to a candidate. If we are interested in how much money the average donor gave, then we must have a way to link that donor's transactions together. In our record linkage for the 2007-2008 election cycle, the mean donation to federal candidates, parties or political action committees was $949. This is the number that is usually reported by academics. However, if donations are linked and aggregated, the mean amount donated by a contributor within the election cycle turns out to be $1,307. Thus, record linkage can make a big difference, even in simple summary statistics.

Government agencies rely on campaign organizations to report data, and there are sometimes slight differences across transaction records. Consider for example Table 1, which contains names and street addresses taken from state and federal donations in Colorado in the 2008 election cycle.

While it is not difficult to see which donations come from the same person

Table 1: Sample Records from State and Federal Donations in Colorado for the 2008 Election Cycle

Name	**Street Address**
Johnson, Mark	Pfizer, Inc.
Johnson, Mark R.	Pfizer, Inc.
Johnson, Mark	640 Fairfield Ln
Johnson, Mark	640 FAIRCHILD LANE
Johnson, Mark	16 Vista Rd
Johnson, Mark K.	16 Vista Road
Johnson, Mark	328 Sutherland Place
Johnson, Mark S.	328 SUTHERLAND PL.
S. Johnson, Mark	328 SUTHERLAND PL
Mark, Johnson	328 SUTHERLAND PL.
Richardson, Mark Johnson	10025 S Blackbird Pl

by eye, it is more difficult to do this over millions of records by machine. A database join requires exact matching, and the difference in middle initials (or even whether a street name is abbreviated) makes this difficult. Besides the slight differences in middle initials, one campaign reversed the last name and first name in its reports. Furthermore, even in the donations listed here, it could be that the Mark Johnson at 640 Fairfield Ln (a residential address) is the same as the Mark (R.) Johnson at Pfizer, Inc. (a business address), since the residential address is in a suburb of Denver, and the Pfizer, Inc. address is in the city limits of Denver.

In donations records, we usually only have names and addresses to match on. Of course, we also have information on the campaign donated to, as well as the amount, so that we could, for example, designate candidate party as a matching field. However, we choose not to match on campaign information because we wish to explain any partisan consistencies rather than assume them.

To the best of our knowledge, very few academics have attempted to address the problem of record linkage over campaign donation databases. We are aware of the *PoliMatch* software, originally developed by Polimetrix. We do not know whether Polimetrix (since acquired by YouGov) continues their work on *PoliMatch*. From a recent list of summer research projects, it

appears that Jonathan Wand may be working on this at Stanford.[1]

One can, of course, buy expensive record linkage software off the shelf. However, such packages are generally not tailored for the limited information available in campaign donation databases. Different "good government" organizations have taken FEC records and done some linkage, often making the resulting information available. Fundrace.org (now hosted at huffingtonpost.com) has taken the candidate electronic daily reports, and connected them to Google Maps, so that you can look up donors by zip code or look at donors in your neighborhood geographically. However, this lists transactions singly rather than attempting aggregation. For example, Paul Rogers, who gave two donations, is listed as two separate individuals, one at 524 Vintage Drive and one at 524 W Vintage Dr.

The Center for Responsive Politics, in its opensecrets.org website, provides cleaned data, and also has linked donations to the same individual (and family), particularly for large donors. They use a combination of automated and human examination to determine record linkage. The Campaign Finance Institute also appears to use a combination of automated and human record linkage. Because we are interested in linking millions of records, human record linkage is not plausible. Furthermore, computationally, it is not plausible to compare every record to every other record (details below).

In the next sections, we explain our methodology of record linkage, and compare its results to a human linked database. Then we examine what difference using a linked database makes in the results. Finally, we discuss other applications and our future plans.

2. RECORD LINKAGE

Record linkage, also known as duplicate record detection, identity resolution, deduplication, and coreference resolution, consists of discovering matching records within a data collection, or combining multiple overlapping data collections, such that records that are believed to refer to the same entity are indeed treated as a single entity. Research in record linkage has its origins in the work of Newcombe and colleagues, who devised a probabilistic matching mechanism, based on sophisticated, hand-crafted comparison rules (Newcombe et al. 1959). That work was later formalized by Fellegi and Sunter who provided a formal framework, which remains the basis of most modern

[1]See http://politicalscience.stanford.edu/srp.html.

approaches to record linkage (Fellegi and Sunter 1969).

Excellent recent overviews of techniques and research issues relevant to record linkage in general have been compiled by Gu et al. (2003), Winkler (2006), and Elmagarmid et al. (2007). In general, the records that must be linked consist of several fields corresponding to individual pieces of information, such as names, dates and addresses, which are stored as character strings. While it is possible to consider a record as a single string through concatenating its various constituent pieces into one, this generally hinders the matching process. One is better off matching pieces separately and combining the results into a single final decision. Hence, record linkage involves two complementary activities: 1) field matching and 2) record matching. We give a brief overview of each in what follows.

2.1. *Field Matching*

Since individual fields are strings, field matching typically makes use of string metrics to quantify the amount of similarity between field values.[2] The two most common categories of string metrics are phonetic comparison algorithms and pattern comparison algorithms.

Phonetic comparison algorithms rely on how strings, or words, are pronounced to compute similarity. For example, the strings *Christie* and *Kristy* are close under such measures, while the strings *Mark* and *Becky* are less so. It is clear that similarity metrics based on phonetics are language-dependent. Common phonetic algorithms include Soundex (Zobel and Dart 1995), Phonex (Lait and Randell 1993), Phonix (Gadd 1990), and Double-metaphone (Philips 2000).

Pattern comparison algorithms compute similarity based on either the cost of transforming one string into the other, or the number and order of common characters between the two strings. The former types are often called edit-distance algorithms. For example, the strings *Christie* and *Kristy* are not very close under such measures, while the strings *Johnson* and *Monson* are more so. Common pattern comparison algorithms include Levenshtein (1966), Needleman-Wunsch (1970), Monge-Elkan (1996) and Jaro-

[2]Some of these strings can actually be numbers, such as an age or a birth year. In such cases, it is possible, and may even be advantageous, to treat them as such when comparing them. Hence, for example, the difference between two age values could serve as a direct measure of their similarity. We restrict our attention here to the more complex case of non-numeric strings.

Winkler (1995).

Unfortunately, little work has been done in terms of comparing the relative value of these metrics on different types of data. An analysis of performance, restricted to edit-distance algorithms, is presented by Navarro (2001). What is clearly known is that no metric is best for all types of data. Experience does suggest that Monge-Elkan, Jaro-Winkler and Soundex are well suited for name matching (Pfeifer et al. 1996; Cohen et al. 2003). Work on genealogical data also found that a weighted ensemble of these three metrics results in better performance than any of its constituents (Ivie et al. 2007).

Note that the above assumes that fields are atomic in the sense that they contain either a single piece of information, such as a first name or a zip code, or multiple, semantically different pieces of information in a standardized format, such as a full US address of the form [number, street name, city, state, zip code], or a complete name of the form [last name, first name, middle initial]. Hence, standardization is an essential pre-processing aspect of record linkage.

While standardization may often be achieved via simple parsing and disambiguation, as in the case of separating zip codes from state names or abbreviations in a composite address field, there are cases when standardization of non-atomic fields is virtually impossible. As an illustration, consider a situation where a field contains both first name and last name, but the order may vary from one record to another, possibly as a result of discrepancies in data entry. For example, one record contains the name *Boyd George* and the other the name *George Boyd*. In this case, it is impossible to match first names and last names separately with any degree of certainty. Any attempt at disambiguation is prone to error as the syntactically identical names may have opposite semantic meanings. In our example, the name *George* may be both a first name and a last name, making the two composite name fields either identical or completely different. Note that this particular problem may also arise at the record matching level, where there may indeed be two separate name fields, but data entry errors cause their associated semantics to be different across different records.

The result produced by field matching may take the form of either the raw value computed by the selected string metric or a summary value based on thresholds. In general, two thresholds may be defined, one below which we are confident that the two strings are not a match, and one above which we are confident that the two strings are indeed a match. The area between the two thresholds serves as an area of uncertainty. By setting the two thresholds

to the same value, we may force the decision to be crisp. The form of the returned result has an impact on record matching as described below.

2.2. *Record Matching*

Most records consist of multiple fields that may or may not be of the same type. The obvious prerequisite for record matching is that a one-to-one semantic mapping between a meaningful subset of fields of the two records exists—or may be naturally derived. In other words, we must be able to decide what piece of information (i.e., field) in one record corresponds to what other piece of information in the other record.

Once an appropriate mapping has been established, record matching typically proceeds in two stages. In the first one, homologous fields are matched. In the second one, individual field scores are combined into a single, final match score for the record pair.

The matching technique used for each field should be appropriate for the associated field type. A score can then be computed for each pair. It is advisable that all returned scores are of the same nature (i.e., either raw or threshold-based) as this simplifies combination. If the returned scores are raw values, it may also be necessary to normalize them so that no single field carries more weight than another only on the basis of the range of values of its selected metric. For example, if the metric applied to field A returns values in the range [0,1] while the metric applied to field B returns values in the range [0,100], differences in B might have more impact on the overall record similarity than differences in A.

Once individual field scores have been computed for all shared fields, many combination approaches are possible, from relatively simple ones based on, for example, Jaccard similarity or average similarity between fields, to more advanced ones using machine learning algorithms (Bilenko et al. 2003). If, as pointed out above the mapping is correct, but errors are likely, one can use a kind of multiset approach where the "offending" fields of one record are matched against the "offending" fields of the other in a pairwise fashion, and individual scores are combined. We show an example of one such solution in our approach to matching FEC data.

2.3. *FEC Record Matching*

We perform our record matching procedure on the daily reports available through the FEC's FTP site,[3] rather than the more generally used "cleaned up" data provided by the FEC. The relevant fields or attributes for linkage are name, zip code and street address. Titles such as Sr. and Mrs., as well as all punctuation marks, were removed from the name field.

Each Individual can have a last name, a first name and a middle name. However, there exist inconsistencies in the ordering of name components. Therefore, we use a kind of multiset approach to account for possible misalignments as follows.

Assume our data contains two individuals A and B whose names have been recorded as:

	A	**B**
First Name	Jason	Anderson
Last Name	Anderson	Johnson
Middle Initial	S	T

We begin by building all possible combinations of name components and rank them in descending order of their matching scores using the Jaro-Winkler metric:

A Component	**B Component**	**Score**
Anderson	Anderson	1.0
Jason	Johnson	0.73
Anderson	Johnson	0.69
Jason	Anderson	0.0
Jason	T	0.0
Anderson	T	0.0
S	Anderson	0.0
S	Johnson	0.0
S	T	0.0

Starting from the top of the list, we consider each pair in turn and select it provided that the names it contains have not been used in a previous selection. Continuing with our example, we would select the first pair (*Anderson*, *Anderson*) and the second pair (*Jason*, *Johnson*). We would then leave out

[3]See http://www.fec.gov/finance/disclosure/ftpdet.shtml.

the next 6 pairs as they each contain at least one name that was part of an earlier selection. Finally, we would select the last pair (S, T). Thus, after alignment, the name A_{name} =*Anderson, Jason S* for A can be compared with the name B_{name} =*Anderson, Johnson T* for B.

While this works well in general, there are, of course, a few situations where this approach fails. Consider again, a record containing the name *George Boyd* and another the name *Boyd George*. The above approach would line these two names together (*George*, *George*) and (*Boyd*, *Boyd*) so that they would be deemed the same individual, when they may indeed be two different persons. There is, however, no way to avoid such difficulties. Either one trusts the file format which may cause true alignments to be missed, or one uses the above approach which may cause false alignments to be created. We feel that the former is riskier in this context than the latter, and thus proceed with our multiset approach.

Addresses are also compared using the Jaro-Winkler metric, and the overall matching score, M_score(A, B), between A and B is then computed as follows.

$$l^i_{name} = \text{len}(A^i_{name}) + \text{len}(B^i_{name})$$
$$l_{addr} = \text{len}(A_{addr}) + \text{len}(B_{addr})$$
$$\text{sc}_{JW}(A, B) = \frac{\sum_{i=1}^{n} l^i_{name}\text{JW}(A^i_{name}, B^i_{name}) + l_{addr}\text{JW}(A_{addr}, B_{addr})}{\sum_{i=1}^{n} l^i_{name} + l_{addr}}$$

$$\text{M\_score}(A, B) = \text{sc}_{JW}(A, B) - \text{ZipPenalty}(A, B)$$

where A^i_{name} (respectively, B^i_{name}) is the ith name component of A (respectively, B) as determined by the alignment procedure described above, $\text{JW}(x, y)$ is the Jaro-Winkler matching score of the strings x and y, n is the number of name components, and $\text{len}(x)$ is the number of characters in the string x.

The term ZipPenalty(A, B) is based on the map distance between zip code areas and is computed as follows. Let ZipA and ZipB be the zip codes of individuals A and B, respectively. Using a specialized look-up table, we retrieve the set of coordinates (longitude and latitude) associated with ZipA and ZipB. We then compute the Euclidean distance between them:

$$\text{dist(ZipA,ZipB)} = \sqrt{(\text{ZipA}_{long} - \text{ZipB}_{long})^2 + (\text{ZipA}_{lat} - \text{ZipB}_{lat})^2}$$

and finally define the penalty term for A and B as:

$$\text{ZipPenalty}(A, B) = w.\text{dist(ZipA,ZipB)}$$

where w weighs the penalty on the overall matching score. The closer two zip codes are the smaller the penalty is, while the farther two zip codes are the larger the penalty. Hence, if two individuals appear very similar, but their addresses are actually geographically far apart, the overall similarity score between them is reduced. While there may be cases where an individual commutes across large distances for work purposes or possesses several residences spread over a large area, this will not be true of most people. ZipPenalty thus offers a simple mechanism to avoid overlinkage when linking across wide areas such as the entire United States.[4] We return to this issue in the next section. Empirically, $w = 0.002$ was found to give good results.

To decide whether two individuals are the same, we threshold the raw matching score to obtain the following simple decision function:

$$\text{match}(A, B) = \begin{cases} 1, & \text{if M\_score}(A, B) \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

In our work here on donation record linkage, empirical results suggest that $\theta = 0.88$ achieves good performance.

2.4. *Nationwide Record Linkage*

Record linkage is by nature a very slow process. Given a collection of records to link, the naive approach would be to take every record in the collection and compare it with every other record. While this approach guarantees that no possible match will be overlooked, it is computationally prohibitive as the amount of time it requires is quadratic in the number of records in the collection. Hence, if we wish to link 10 million records —the number of records contained in the FEC database in the 2007-2008 election cycle— no less than 100 trillion comparisons are necessary. This number of comparisons would take most modern computers, except for the fastest machines in the world, from weeks to months to carry out.

[4]While always of some value, the penalty term is particularly useful when street addresses are missing from the records, or omitted from the computation (e.g., due to lack of standardization or other related problems).

Another problem when linking large numbers of records is what may be viewed as probabilistic overlinkage. Consider for example, the two names *Bob Smith* and *Bobby Smith*. Both of them could be the same person, and therefore would be linked together. Assume now that a second person also named *Bobby Smith* shows up in one of the other records. It appears that this second person is the same as the first. Yet another record appears with the name *Bobby Smithers* and it is determined that its record should be linked with that second *Bobby Smith*'s. And eventually, the second *Bobby Smith* and the first are also linked. We have now linked *Bob Smith*, *Bobby Smith* and *Bobby Smithers* together. This overlinking problem gets worse as the number of records in the collection grows, so that the amount of linkage error grows as the number of records to match grows.

Hence, some mechanism is required to reduce the number of comparisons while not adversely affecting the accuracy of the linkage process too much. Several solutions have been proposed, including blocking, sorted neighborhood, clustering, canopies and set joins (Bilenko et al. 2003). As a first attempt, we used a canopies-like approach in which we divided the nation into overlapping units centered on each individual and extending to some pre-defined distance from it in all directions. While natural, this approach led to several problems due to the overlaps. In particular, provided A's area overlaps with B's and B's area overlap with C's, we would first link each area independently of each other and then run into the problem of having to join the results together as there are people in both A's and B's area and people in both B's and C's area. Furthermore, the overlap induced a transitivity problem, similar to the probabilistic overlinkage described above, as one might envisage a situation where two records link between A and B, and the record in B also links with one in C, thus creating a link between the individual in A and the individual in C.[5] Taken to the limit, all individuals with the same common name, e.g., *John Smith*, in the nation are linked together. This is clearly unacceptable.

What we needed was a kind of blocking approach, where we could divide the nation still, but do so in a non-overlapping fashion, while not hindering linkage. As it turns out, the United States' Office of Management and Budget has divided the nation in non-overlapping areas known as Metropolitan

[5]Note that here the penalty term does not help as each link takes place in a small area and the transitive link is not computed explicitly thus essentially factoring out the distance between A and C.

and Micropolitan Statistical Areas (MSAs). Incidentally, and conveniently for our purposes, each MSA can be considered as an area independent of other areas used for statistical purposes because usually people move and do business within their MSA, and rarely move out of their MSA or live and work across MSA boundaries. Our solution to link across the entire United States is therefore based on first blocking on MSA (i.e., assigning each record to its associated MSA based on the zip code in the address field) and then performing record linkage in each MSA independently. Within this context, the use of our penalty term seems superfluous. However, some MSAs are rather large (e.g., less dense areas of the nation), so we choose to retain it in the formula to account for possible overlinkage within these.

In addition to reducing the size of the collections over which linkage has to be performed, our blocking procedure also enables a parallel implementation, where each MSA can be linked on a different computer (or CPU in a supercomputer environment). Hence, the time it takes to link the entire nation is the same as the time it takes to link the largest MSA. While we did not have sufficiently many machines to farm out each MSA to a separate one at the same time, we used 4 to 5 standard PCs and kept cycling MSAs through them as the previous ones would complete. We downloaded the shapefile from the Census's website,[6] which contains all 369 Metropolitan and 578 Micropolitan areas. Zipcodes were assigned to the MSA they were most geographically proximate to, using simple Euclidian distance in ArcGIS, and all individuals in the zip code were then assigned to the corresponding MSA. The FEC records are spread over these MSAs such that the largest MSA (Washington-Arlington-Alexandria) contains over 625,000 records, the smallest MSA (Guyama, Puerto Rico) contains less 12 records, and the average size of MSAs is about 9,790 records.

3. LINKAGE VALIDATION

To validate our approach and determine linkage accuracy, we tested it against two different benchmarks. In the first, we use hand-labeled data to compare the linkages established by our automatic approach to those advocated by our human annotators. In the second, we use self-reported donation information from a random sample of donors and compare it with what our approach suggests these donors would have donated.

[6]Available at: http://www.census.gov/geo/www/cob/mmsa2003.html#ascii.

3.1. *Agreement with Manual Linkage*

A small portion of the campaign contributions were selected to be manually linked by humans. The areas that were selected for manual linkage were portions of the states of New York, Nevada, and Utah. In total, approximately 7,500 donations were manually linked. These same donations were then linked by computer using the above explained process. Exact duplicates, that is, records that are in every detail identical, were removed prior to record linkage.

The result of linkage can be viewed as a partition of the database into a set of clusters, where each cluster is a group of records that have been deemed to represent the same person. It is then possible to examine records in a pairwise fashion, to determine whether they appear in the same manually-generated and computer-generated clusters. For purposes of comparison, we assume that the manual linkage was completed without any errors, and that any deviation between the two clustering results must be due to an error in computer linkage.

Note that the manual labeling was performed prior to our computerized record linkage work, at a time when we were not aware of the FEC's daily electronic file uploads. The labelers were presented records from the cleaned data on the FEC website, which does not contain street addresses. Hence, the linkage is done here almost exclusively on names, except for the small implicit address bias induced by the ZipPenalty term.[7] Consequently, our results may be slight underestimates of the actual performance of our proposed approach.

In our pairwise analysis, each record is paired up with every other record exactly once, and each resulting pair is then accounted for as follows.

[7] That bias is small because the records are all localized to individual states.

a: Number of pairs whose elements are in the same cluster in both the manually-linked and the computer-linked data. This is the number of correct matches (or true positive), i.e., records that should have been linked and were.

b: Number of pairs whose elements are in different clusters in both the manually-linked and the computer-linked data. This is the number of correct mismatches (or true negative), i.e., records that should not have been linked and were not.

c: Number of pairs whose elements are in the same cluster in the computer-linked data but in different clusters in the manually-linked data. This is the number of incorrect matches (or false positive), i.e., records that should not have been linked but were.

d: Number of pairs whose elements are in the same cluster in the manually-linked data but in different clusters in the computer-linked data. This is the number of incorrect mismatches (or false negative), i.e., records that should have been linked but were not.

As there is no consensus as to which metric is best for measuring the quality of clustering, we use the above quantities to compute a number of widely used statistics that, taken together, provide a strong sense of the overall quality of the computer-generated linkage with respect to the manual linkage. In particular, we consider:

- Precision: The ratio of correct matches to the total number of actual matches.
$$P = \frac{a}{a+c}$$
Precision ranges in $[0, 1]$. Higher values of precision indicate that the computer is linking most of the records it should.

- Recall: The ratio of correct matches to the total number of computed matches.
$$R = \frac{a}{a+d}$$
Recall ranges in $[0, 1]$. Higher values of recall indicate that the computer is not linking too many of the records that it should not.

- F-score: The geometric mean of precision and recall; an attempt at combining both metrics into a single one to account for the natural

trade-offs between them.

$$F = \frac{2 \times P \times R}{P + R}$$

The F-score ranges in $[0, 1]$. Higher F-score values are achieved as both precision and recall are high.

- Rand Index: A measure of the amount of agreement between the manual and the computer linkages (Rand 1971). It may be viewed as a measure of the accuracy of the linkage.

$$RI = \frac{a + b}{a + b + c + d}$$

The Rand index ranges in $[0, 1]$. Higher values indicate stronger agreement between the computed linkage and the target linkage.

- Adjusted Rand Index: An extension of the Rand index proposed by Hubert and Arabie (1985) to compensate for records that may have been linked by chance.

$$ARI = \frac{2(ab - cd)}{(a + c)(c + b) + (a + d)(d + b)}$$

Tables 2-4 summarize the relationship between manually-linked and computer-linked records for New York, Nevada and Utah, respectively.

Table 2: Computer vs. Manual Linkages: New York

		Computer	
		Linked	**Not Linked**
Manual	**Linked**	87,151 (0.91%)	6,727 (0.07%)
	Not Linked	8,176 (0.09%)	9,501,099 (98.93%)

In all cases, the linkage quality metrics are rather high as shown in Table 5. The last row corresponds to the overall linkage quality when all 3 manually-labeled samples are aggregated.

The high values of the Rand index and the adjusted Rand index suggest that there is strong agreement between the computer-generated linkages and

Table 3: Computer vs. Manual Linkages: Nevada

		Computer	
		Linked	**Not Linked**
Manual	**Linked**	29,986 (1.12%)	4,355 (0.16%)
	Not Linked	12,518 (0.47%)	2,631,596 (98.25%)

Table 4: Computer vs. Manual Linkages: Utah

		Computer	
		Linked	**Not Linked**
Manual	**Linked**	11,384 (2.25%)	1,580 (0.31%)
	Not Linked	1,320 (0.26%)	492,237 (97.18%)

the manually-labeled records. Precision is also rather high showing that our approach misses very few of the actual linkages. Similarly, recall, except in the case of Nevada where the value is a little lower, has relatively high value confirming that our approach successfully avoids overlinking.

Interestingly, although we assumed that the manually-linked clusters were correct, there is some evidence that occasionally the computer-linked clusters are actually more accurate than the manually-linked clusters. For example, consider the following two pairs of donations, where occupation is also shown.

Schwartz, Bernard L. Mr.	New York	NY	10021	Loral Corporation
Schwartz, B. L	New York	NY	10021	Loral Space Communications
NELDICH, DAN	NEW YORK	NY	10028	GOLDMAN SACHS
Neidich, Dan	New York	NY	10028	Goldman Sachs/Managing Partner

Both of these pairs of donations were put in separate clusters by the manual labelers, but they were clustered together when linked by the computer. Upon further examination, it seems clear that the computer's decision is actually the correct one in these instances.

On the other hand, there are still a few cases where the computer misses some matching records. For example, the following pairs of donations, matched by the manual labelers, were not clustered by the computer, when it appears that indeed they should have been.

Table 5: Cluster Quality

	P	R	F	RI	ARI
New York	0.93	0.91	0.92	0.998	0.920
Nevada	0.87	0.71	0.79	0.993	0.777
Utah	0.88	0.90	0.89	0.994	0.884
Overall	**0.91**	**0.85**	**0.88**	**0.997**	**0.880**

Taylor, Margaretta Ms.	New York	NY	10022	Homemaker
Ms. Margaretta Taylor	New York	NY	10022	Homemaker
NEIDICH, BROOKE GARBER	NEW YORK	NY	10028	HOMEMAKER
Neidich, Brooke	New York	NY	10028	Homemaker

In the case of the second pair, the score may have been reduced due to the presence of the extra middle name in one of the records. However, for the first pair, we would have expected our multiset approach to restore the correct alignment and thus produce a high similarity score.

Similarly, there are a few instances where the computer links records that should not be. For example, the following donations were matched by the computer, when it is clear that, as suggested by the manual labelers, they should not be.

PATRICOF, ALAN J	NEW YORK	NY	10021	APAX PARTNERS
PATRICOF, SUSAN	NEW YORK	NY	10021	HOMEMAKER

In this case, the similarity in first names is likely the cause of the computer's mistake. In the following example, however, it would appear that while the labelers marked the two records as different, the computer's linking may be correct. Having access to the street address would help resolve the problem, as the labelers' decision may be due to a misspelling of the first names.

HURST, FEM K	NEW YORK	NY	10128	
HURST, FERN	NEW YORK	NY	10128	RETIRED

Overall, the quantitative results as well as the above sample of qualitative findings lend credibility to our proposed automated record linkage approach and strongly suggest that it is rather effective at avoiding both overlinking and underlinking.

3.2. *Agreement with Self-reported Information*

In early 2009, we also used the results of our linkage of the 2007-2008 campaign finance records to draw a representative sample of itemized contributors to federal candidates. Previous studies of campaign contributors have relied on the disaggregated contributions in their original samples. These studies will generally attempt to rectify the obvious problems this creates by hand-matching each name in their sample to determine how often the individual had given in the past. This post-hoc weighting method has obvious drawbacks, and it would be preferable to sample individuals directly as we are able to do with the linked database.

After drawing the sample, we administered a survey to these individuals. In the survey, we asked several questions pertaining to their contribution behavior that are (in theory at least) objectively verifiable through the information we collected in the match. As of this writing, questionnaires continue to be returned, and the results presented here are based on 1,464 returns from individuals whose name and address information we collected from FEC records.[8] These individuals either filled out an online or paper questionnaire. At one point in the survey, individuals were asked to indicate which of the major presidential candidates they contributed to at any point during the 2007-2008 election cycle. By comparing their self-reported contribution behavior with what we observe in the linked database, we can test the reliability of the matching procedure.

Comparisons between observed behavior (from the linkage) and self-reported behavior (from the survey) are complicated somewhat by the FEC reporting requirements. Because contributions are not disclosed until they reach the $200 threshold, we never observe the behavior of individuals that contribute to a candidate below this amount, so it is possible that self-reported and observed behavior will not match for individuals who give near the threshold. For example, an individual who was disclosed to the FEC for a three $75 donations to Obama and who was not disclosed for two $75 donations to Biden would come up as a false positive for Biden in our records (the individual's self-report would not match with the information we had in the linked database). For this reason, it is perhaps most instructive to examine

[8]The survey included individuals whose contact information was collected from other sources as well. In order to keep the comparisons valid, we only report the results from the itemized FEC database here.

the true positive rate (or conversely the false negative rate). If the linkage were successful (and individuals accurately reported their own behavior), the false positive rate should be zero. Table 6 shows the true positive rate for the major presidential candidates in our sample.[9]

Table 6: Results for Major Presidential Candidates

	False Positive	**True Positive**
Biden	1	5
Clinton	16	107
Edwards	8	26
Giuliani	9	30
Huckabee	0	22
McCain	40	334
Obama	27	455
Paul	3	49
Richardson	3	12
Romney	8	64
Total	115	1,104
	9.6%	90.4%

A closer examination of the data reveals that the greatest portion of the false negatives arises from the individuals in the sample who claimed not to have contributed to any candidates. Excluding these individuals improves the true positive rate to 97.2%. It is possible that these individuals were especially sensitive about their privacy and chose not to reveal their contributions even with the anonymity assurances we gave.

[9]One plausible explanation for the discrepancies between observed and self-reported behavior is the possibility that the individual who completed the survey is not the same as the individual to whom the survey was addressed (if, for example, the spouse of the intended recipient filled it out). We were fortunate to have the cooperation of Catalist (a microtargeting firm that has extensive demographic information culled from voter files and consumer databases). Rerunning the analysis in Table 6 with the suspicious cases removed (individuals who do not match in terms of their gender, age, or race between the self-reported demographic information we collected and the demographic information available in the Catalist database) results in a slight improvement to the true positive rate (from 90.4% to 92.3%).

4. ANALYSIS OF CAMPAIGN CONTRIBUTORS

Perhaps the best test of our new linkage method can be found in its practical application. Political analysts and journalists often report descriptive statistics about donations and donors, such as the average donation in a reporting cycle.[10] Such statistics are, of course, greatly impacted by the choice of unit of analysis, i.e., donation or donor. Lacking an effective and accurate way of linking donation records, most researchers are confined to using donation as the unit of analysis, which in turn affects the conclusions being reached. Using our record linkage method, we highlight some significant differences in the results when one considers donors, rather than donations, as the unit of analysis.

Our data comes from two complementary sources, as follows.

1. *FEC Records.* To appear in the FEC records, individuals must donate at least $200 in the aggregate to any one candidate for federal office. The burden of disclosure is on the candidate, who is responsible for tracking (and aggregating) the contributions made to his/her campaign. Individual contribution limit for the 2007-2008 cycle was $2,300 for each candidate-election. For example, individuals were permitted to donate $2,300 to Obama for the primary and $2,300 for the general election. McCain took the public financing grant for the general election, and consequently, individuals were not permitted to donate to his campaign for the general election. However, both major party candidates established joint party-candidate "victory" funds, allowing individuals to donate beyond the $2,300 limit, up to the maximum allowable amount of $28,500.[11]

2. *Campaign-specific Records (CSR).* The Obama and McCain campaigns are generally thought to have pursued different kinds of strategies, particularly regarding "small donors" (i.e., less than $200 total donations).

[10] For example, in discussing the second quarter fundraising statistics of 2007, *The Washington Post* reported that, "The vast majority of Obama's donors gave in relatively small amounts....The average donation was $202." (Solomon 2007)

[11] The FEC records contain "negative" donations, corresponding to donations returned to individual donors for a variety of reasons (e.g., contribution limit exceeded). The number of such entries is relatively small (less than 1% of the number of donations), so we simply excluded them from our analyses.

In an attempt at discovering whether there were indeed different patterns in small donors between the two campaigns, we also use random samples of small donors generously supplied by the Obama (10,000 of 3.2 million reported small donors) and McCain (7,600 of 613,385 reported small donors) campaigns.[12]

Table 7 shows aggregate summary statistics for the publicly available FEC donations, as well as individual summary statistics for Obama-only and McCain-only donors, based on the FEC data augmented by the CSR data. The small donor samples are weighted by factors of 80.6 and 322.1 for McCain and Obama donors, respectively, to reflect the numbers of small donors reported to us by each campaign. In all three cases, values in the first row are obtained using the donation as the unit of analysis, i.e., using unlinked records, while values in the second row are obtained using the donor as the unit of analysis, i.e., using linked records.

Table 7: Donation/Donor Summary Statistics (in Dollars) for 2007-2008

	Mean	**Median**	**Std. Dev.**
Overall			
By Donation	949	500	2,010
By Donor	1,307	500	3,793
Obama (with weighted small donors)			
By Donation	71	28	451
By Donor	104	50	546
McCain (with weighted small donors)			
By Donation	199	38	1,259
By Donor	269	61	1,457

In the publicly available records, the mean donation was $949, whereas the mean amount given by a donor was $1,307, about 38% higher. A similar observation can be made on the specific campaigns, with 48% and 35% higher values for Obama and McCain, respectively. Furthermore, the candidate-specific data suggests that Obama seems to have attracted more repeat small donors than McCain did. We take a closer look at the differences between

[12]We had to sign appropriate non-disclosure agreements to obtain this data.

the two campaigns in the following. We restrict this analysis to the publicly available FEC data.

As reported by the media, McCain received more of his money from larger donations (unlinked) than Obama. To qualify this assertion, we first show the distribution of donations and donors to the Obama and McCain campaigns in Table 8 by amount.

Table 8: Percentage Distribution of Obama and McCain's Donations and Donors by Amount

	Unlinked		**Linked**	
Amount	**Obama**	**McCain**	**Obama**	**McCain**
200-500	30	12	21	13
500-1,000	11	9	19	13
1,000-2,300	19	20	22	25
2,300-4,600	21	26	23	35
4,600-28,700	17	23	14	11
28,700-56,400	3	6	1	1
Over 56,400	0	4	0	1

These statistics show that about 59% of McCain's donations were in amounts of $2,300 or more. This compares to only 41% for Obama. Similarly, Obama received 30% of his donations in amounts between $200 and $500, while McCain received only 12% of his donations in amounts of the same size. However, once we apply record linkage and the multiple donations by a single donor are aggregated, the differences are not as large. About 48% of McCain's contributions and 38% of Obama's contributions came from donors who gave $2,300 or more. Similarly, 13% of McCain's contributions and 21% of Obama's contributions came from donors who gave between $200 and $500. While it is clear that Obama's donors were generally smaller, the difference between the McCain campaign and the Obama campaign is not as stark once the donations are linked.

Table 9 further (and maybe more directly) addresses the question of how the linkage affects the way we think about the distribution of Obama and McCain donors. Whereas Table 8 is concerned with the number of donations and donors, Table 9 focuses on dollar amounts raised.

Table 9: Percentage Distribution of Obama and McCain's Total Itemized Individual Donations Raised for Donations and Donors of Different Sizes

	Unlinked		**Linked**	
Amount	**Obama**	**McCain**	**Obama**	**McCain**
200-500	22	12	9	9
500-1,000	21	13	16	13
1,000-2,300	25	22	24	21
2,300-4,600	20	25	22	20
4,600-28,700	11	18	23	26
28,700-56,400	2	6	4	6
Over 56,400	0	3	2	5

These statistics show that if we were to consider only the unlinked records, we would come to the conclusion that Obama raised 22% of his (itemized) money from donations between $200 and $500, against only 12% (half as much) for McCain. However, when we link donations, we see that Obama only raised 9% of his money from donors in this category, which is the same as McCain's 9%. The graphs in Figures 1-4[13] provide another view of these effects. They are histograms of the distributions of donations vs. donors for small and large donation amounts, for both Obama and McCain. The horizontal axis is the individual donation or aggregate donor amounts and the vertical axis is the square root of the frequency. We use the square root transformation to magnify the right hand side of our graphs. In addition to clearly showing that the linkage causes the distribution to shift to the right, as expected from the results in Table 7, these graphs also show that the linkage significantly changes the distribution of the sources of Obama's campaign funds at smaller levels, but has less of an effect for McCain. This suggests that Obama was much more likely to receive multiple smaller donations. Most of the movement in the McCain graph happens among donors giving $2,300 and then giving more.

We provide several other comparisons in Table 10. This table contains information on publicly available donations to all political committees, including congressional campaigns, and political action committees. As with

[13]Figures 1 and 2 are based on the CSR data

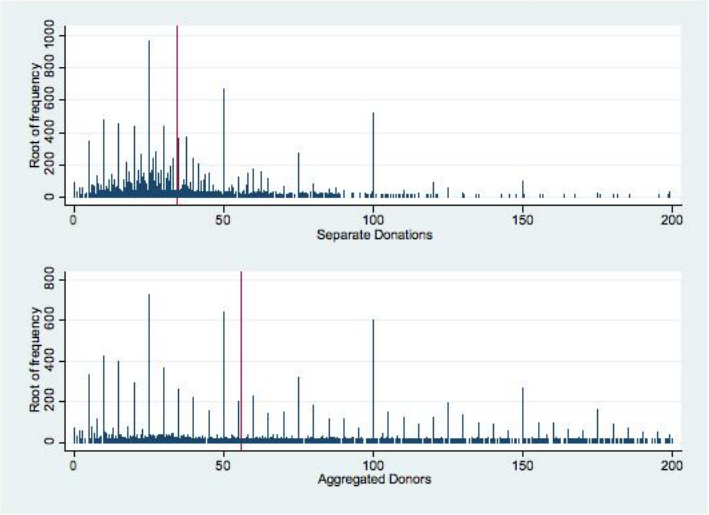

Figure 1: Small Donations vs. Donors for Obama

the previous tables, we examine different percentages by donation or donor amount. The first number in a cell is the number or percentage when the donations are linked, i.e. the unit of analysis is the donor. The number in parentheses directly below is the equivalent number or percentage when the unit of analysis is the donation, i.e. unlinked.

The first (n) column reports the number of donors (donations). The second (%Obama) and third (%McCain) columns show the percentage of the total number of donations that were made to, respectively, the percentage of the total number of donors who gave to, the Obama and McCain campaigns. For example, 8% of all contributions between $200 and $500 went to McCain, and 5% of all donors who gave in that range donated to McCain. The fourth (#Don.) column counts the mean number of distinct contributions made by donors. The fifth (%House) column examines what percentage of donors (donations) contributed to House candidates, as opposed to Senate, presidential, PACs, and parties. Finally, the last (%Cand.) column shows what percentage of donors (donations) contributed to candidates, rather than parties or PACs.

These statistics again show the clear impact that linkage has on the conclusions one may reach about donations, donors and campaign results. The following are a few observations based on Table 10.

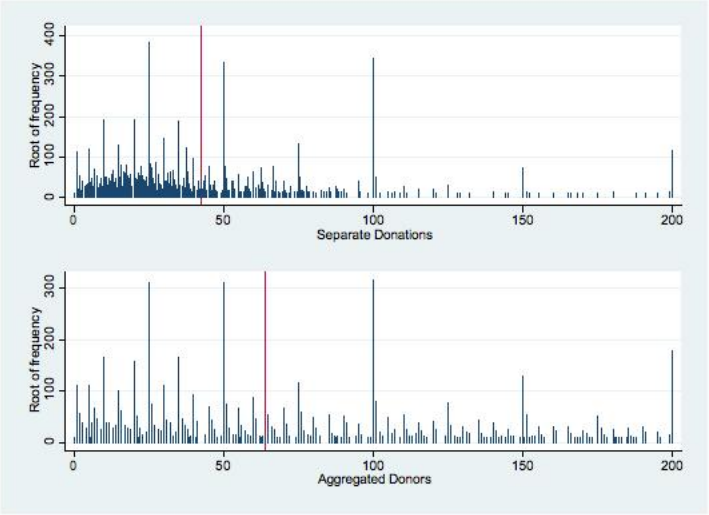

Figure 2: Small Donations vs. Donors for McCain

- While there were over 8 million separate donations, there are only about 2 million donors.
- When we compare the percentages of all (publicly disclosed) donors contributing to the Obama and McCain campaigns (second and third column), the conventional wisdom is again confirmed that Obama received more money from smaller donors (and donations) than McCain did.
- Among smaller donors (contributions between $200 and $500), the mean number of contributions is about 2.5. In other words, the average smaller donor contributed to one campaign between 2 and 3 times. This is in contrast to larger donors, who gave to more different campaigns, and gave more donations overall. While media and the campaigns have often emphasized how smaller donors were giving multiple donations, it is the larger donors that are giving more frequently to multiple campaigns multiple times.[14]
- While large donations are rarely given to House candidates, large donors

[14]Note that while this is clearly true, it may be a little misleading. Once an individual gets to amounts in excess of $2,300 (the legal limit) they are necessarily giving multiple times or to multiple campaigns.

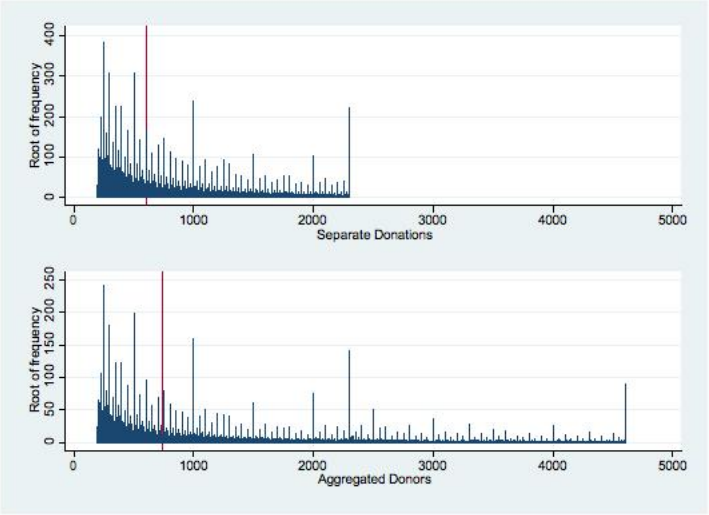

Figure 3: Large Donations vs. Donors for Obama

often give to House candidates. This implies that large donors are giving to other candidates in larger amounts.

- When considering the impact of linkage on the fifth (%Cand.) column, we observe very little difference between donations and donors, except in one category. Thus, failing to link records in this instance would not lead to major differences in conclusions: as contributions increase, donors are more likely to contribute to candidates rather than PACs and parties. The one exception to this general pattern is found in the 4,600-28,700 row. This squares with the contribution limits that were in place for the 2007-2008 election cycle, as individuals were not permitted to contribute to candidate committees in amounts larger than $2,300 at a time. However, individuals could give in larger amounts to PACs ($5,000 per year) and local party committees ($10,000 per year). In this table the joint victory committees were included as candidate donations accounting for the 45% figure reported. When we aggregate the donations across, we see that the overwhelming majority of individuals who make these larger donations also give to candidate committees.

The results reported in this section further demonstrate the validity of our approach, and clearly highlight the importance of accurate record linkage to

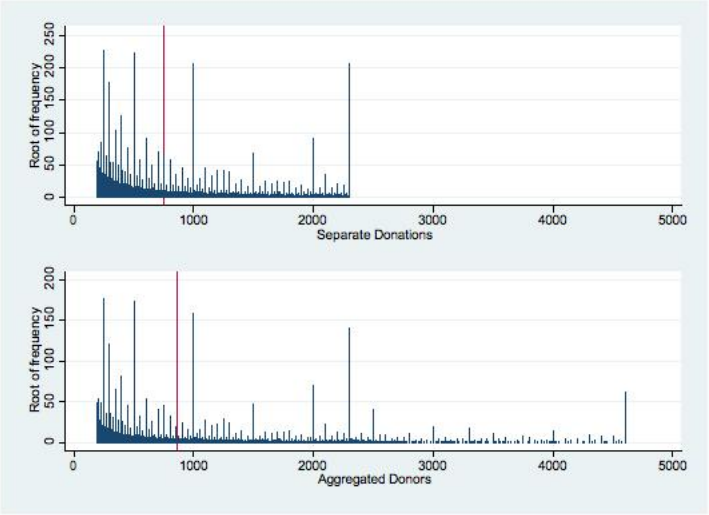

Figure 4: Large Donations vs. Donors for McCain

substantiate claims made about campaign activities and results, when donors are to be taken as the unit of analysis.

5. CONCLUSION

In this paper, we have briefly described the technique of record linkage and shown how it can be implemented effectively in the context of nationwide donation data. By using a kind of blocking approach based on the US Office of Management and Budget's well-defined Metropolitan and Micropolitan Statistical Areas, we are able to parallelize our approach, thus making nationwide linkage over millions of records feasible.

We applied our technique to data from the 2007-2008 election cycle, both in validation and in generalization. We validated our approach by comparing its performance against that of human labelers as well as against results obtained from self-reported information. We generalized it by taking a fresh look at the 2007-2008 election data, deriving global statistics as well as statistics related to the Obama and McCain campaigns. We were able to show the clear impact of linkage, provide scientific confirmation to conventional wisdom, and derive new insight about these campaigns.

In addition to linking data within the 2007-2008 election cycle, we have also applied our technique to linking data from the FEC (for national races)

Table 10: Description of Contributors (Contributions) by Amount Given

	n	%Obama	%McCain	#Don.	%House	%Cand.
200-500	1,086,163	25	8	2.5	10	54
	(7,004,312)	(25)	(5)	(-)	(6)	(46)
500-1,000	426,435	32	11	4.7	17	72
	(542,110)	(21)	(11)	(-)	(27)	(72)
1,000-2,300	308,601	29	16	5.4	27	82
	(465,214)	(19)	(13)	(-)	(29)	(79)
2,300-4,600	147,821	29	19	5.7	31	92
	(223,002)	(22)	(18)	(-)	(22)	(90)
4,600-28,700	86,009	33	26	8.1	50	91
	(52,548)	(18)	(17)	(-)	(3)	(45)
28,700-57,400	4,967	47	37	13.1	64	98
	(1,352)	(35)	(45)	(-)	(2)	(83)
Over 57,400	1,716	46	47	21.1	74	98
	(211)	(4)	(84)	(-)	(6)	(94)

data from to state campaign finance records. Furthermore, we plan to add other national databases (e.g. 572 organizations from the IRS database) and link donors across election cycles.

While our results are promising, they also need to be further evaluated. In particular, we would like to compare our linkage against the FEC and CFI linkages. We are satisfied that qualitatively at least, our results are similar. Unfortunately, the FEC's and CFI's methodologies have not been released yet. Hence, a formal quantitative analysis is left as future work.

REFERENCES

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

Cohen, W., P. Ravikumar, and S. Fiendberg (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 73–78.

Elmagarmid, A., P. Ipeitoris, and V. Verykios (2007). Duplicate record detection: A survey. *IEEE Transactions in Knowledge and Data Engineering 19*(1), 1–16.

Fellegi, I. and A. Sunter (1969). A theory for record linkage. *Journal of*

the American Statistical Association 64(328), 1183–1210.

Gadd, T. (1990). PHONIX: The algorithm. *Program: Automated Library and Information Systems 24*(4), 363–366.

Gu, L., R. Baxter, D. Vickers, and C. Rainsford (2003). Record linkage: Current practice and future directions. Technical Report No. 03/83, CSIRO Mathematical and Information Sciences.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Ivie, S., B. Pixton, and C. Giraud-Carrier (2007). Metric-based data mining model for genealogical record linkage. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pp. 538–543.

Jaro, M. (1995). Probabilistic linkage of large public health data file. *Statistics in Medicine 14*(5-7), 491–498.

Lait, A. and B. Randell (1993). An assessment of name matching algorithms. Technical report, Department of Computer Science, University of Newcastle upon Tyne, UK.

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

Monge, A. and C. Elkan (1996). The field-matching problem: Algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 267–270.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys 33*(1), 31–88.

Needleman, S. and C. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*(3), 443–453.

Newcombe, H., J. Kennedy, S. Axford, and A. James (1959). Automatic linkage of vital records. *Science 130*(3381), 954–959.

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

Philips, L. (2000). The double-metaphone search algorithm. *C/C++ Users Journal 18*(6), 38–43.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Solomon, J. (2007). Obama takes lead in money raised. *Washington Post, July 2*, A1.

Winkler, W. (2006). Overview of record linkage and current research directions. Research Report Series (Statistics #2006-2). Available online at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Zobel, J. and P. Dart (1995). Finding approximate matches in large lexicons. *Software: Practice and Experience 1*, 331–345.