Knowl Inf Syst (2015) 45:389–416
DOI 10.1007/s10115-014-0812-5

REGULAR PAPER

Effective record linkage for mining campaign contribution data

C. Giraud-Carrier · J. Goodliffe · B. M. Jones · S. Cueva

Received: 28 February 2014 / Revised: 17 November 2014 / Accepted: 12 December 2014 /
Published online: 20 December 2014

Abstract Up to now, most campaign contribution data 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 a unique repository of donations nor any standard across existing repositories. In order to more meaningfully mine campaign contribution data, political scientists need an accurate way of grouping, or linking, together donations made by the same donor. In this paper, we describe a record linkage technique that is applicable to various sources and across large geographical areas. 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 US election cycle.

Keywords Record linkage · Multiset distance · Domain knowledge · Campaign contributions · Political data

1 Introduction

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. When entities have unique identifiers (e.g., social security number, SKU code), record linkage is, of course, trivial. In many cases, however, no such identifiers exist, and record linkage requires sophisticated matching algorithms.

C. Giraud-Carrier (✉) · S. Cueva
Department of Computer Science, Brigham Young University, Provo, UT 84602, USA
e-mail: cgc@cs.byu.edu

J. Goodliffe · B. M. Jones
Department of Political Science, Brigham Young University, Provo, UT, USA

Record linkage has been used extensively in the medical domain, where patient data from one source must be matched with data from another. For example, in follow-up studies, records of cohort members can be matched with mortality records to ascertain which patients may have died [15]. Similarly, in studies about pregnancy and subsequent child health, earlier records containing information about pregnant women can be linked to later records containing information about their offspring to analyze birth outcomes and find associations between an expectant mother's health behavior (e.g., smoking) and the impact on her child (e.g., cancer) [20,33]. Specialized record linkage systems have also been deployed to support population health research (e.g., see [17,34]). Another important area of application of record linkage is genealogical research, which relies on old data sources that have now been digitized, such as census records and parish registers. For example, in studies of population migration and household composition, records of individuals and families can be linked across several consecutive censuses [9,10,31]. In the context of family history, matching individuals across pedigrees can help to bring together the complementary work of separate researchers [18,29,35]. Finally, there are a number of other applications where record linkage must be used, such as merging administrative and business lists, and identifying duplicates [13,37].

In this paper, we focus our attention on political science. In political science, a primary use for the process of record linkage 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.

Government agencies rely on campaign organizations to report data, and there are sometimes slight differences across transaction records. Consider for example Table 1, which

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

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 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.

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 mining donor-level data, e.g., to know how much money the average donor gave, then we must have a way to link that donor's transactions together. Furthermore, omitting to effectively link donations may lead to inaccurate results. For example, In 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 in the mining of campaign contribution data, even in simple summary statistics.

One of the challenges of donations records is that 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, doing so would potentially bias results. Hence, we choose not to match on campaign information, since this allows to *explain* any partisan consistencies rather than *assume* them. In what follows, we explain our methodology of record linkage and validate it against a human-linked database. We then examine what difference using a linked database makes in the mining of campaign contribution data and report on previously unattainable results about campaign contributors in the 2007–2008 US election cycle. Finally, we show how the same technique can be extended to other types of data to assist researchers.

2 Related work

Research in record linkage has its origins in the work of [25], who devised a probabilistic matching mechanism, based on sophisticated, hand-crafted comparison rules. That work was later formalized by [8] who provided a formal framework, which remains the basis of most modern approaches to record linkage. Excellent recent overviews of techniques and research issues relevant to record linkage in general have been compiled by [4,7,12,38].

To the best of our knowledge, very few researchers 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 (see politicalscience.stanford.edu/srp.html for details).

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 one can look up donors by zip code or look at donors in one's 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 feasible. Furthermore, even computationally, it is not reasonable to compare every record to every other record.

3 Record linkage

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.

3.1 Field matching

Since individual fields are strings, field matching typically makes use of string metrics to quantify the amount of similarity between field values. 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.

3.1.1 Standardization

One can distinguish between two types of fields, atomic and composite. An atomic field is one consisting of a single string, such as a first name or a zip code. A composite field is one consisting of multiple, semantically different strings, such as a full US address or a complete name. Atomic fields can be compared directly. Composite fields, on the other hand, require standardization, before they can be subjected to string matching. Standardization is the process of re-arranging the elements of a composite field so that they all follow a common format. For example, elements of US addresses may be re-arranged into [number, street name, city, state, zip code], and elements of names may be re-arranged into [last name, first name, middle initial].

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 the 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.

3.1.2 String metrics

Once fields have been standardized as needed, their values can be compared using string metrics. The two most common categories of string metrics are phonetic comparison algorithms and pattern comparison algorithms.

Phonetic comparison algorithms compute similarity based on how strings, or words, are pronounced rather than on how they are written. Hence, the strength of a match between two strings depends entirely on how much one string sounds like the other. For example, the strings *Christie* and *Kristy*, and the strings *jeans* and *genes*, though spelt differently, sound the same and thus would be considered very close by phonetic comparison algorithms. On the other hand, the strings *Mark* and *Becky* would be deemed rather different. It is clear that similarity metrics based on phonetics are language-dependent. For example, the strings *mirage* and *garage* would be closer in US English than they would in British English. Common phonetic comparison algorithms include Soundex [39], Phonex [21], Phonix [11] and Double-metaphone [27].

Pattern comparison algorithms, by contrast, consider words in their most basic form, as sequences of characters. They then 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. Under pattern comparison algorithms, the strings *Christie* and *Kristy*, and the strings *you* and *ewe*, would not be very close, while the strings *Johnson* and *Monson* would have a higher degree of similarity. Common pattern comparison algorithms include Levenshtein [22], Needleman–Wunsch [24], Monge–Elkan [23] and Jaro–Winkler [19].

A limited amount of work has been done in terms of comparing the relative value of these metrics. Experience does suggest that Monge–Elkan, Jaro–Winkler and Soundex may be well suited for proper name matching [5,26]. Yet, a rather comprehensive analysis of name matching metrics shows no clear winner [3]. Some studies have even shown that combinations of metrics, via for example weighted ensembles, outperform single metrics [18]. The same is true when considering different data types. Experiments in the context of genealogical record linkage show that different metrics perform best on different types of data, such as names, dates and addresses [28]. Although [3] provides some recommendations in the context of name matching, and [2] offer some guidance in the context of ontology matching, little is known as to which metric, or combination of metrics, to use when. Indeed, it is safe to say that there is no universal string comparison algorithm. Hence, metric selection for field matching is often the result of a mixture of experience and experimentation.

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, a rejection threshold below which one is confident that the two strings are not a match and an acceptance threshold above which one is 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, one may force the decision to be crisp. The form of the returned result has an impact on record matching as described below.

3.2 Record matching

In most practical applications, 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, or field, in one record corresponds to what other piece of information in the other record. For example, if one record had fields *first name*, *last name*, *phone number*, *occupation* and *address*, and the other record had fields *surname*, *given name*, *birthdate* and *address*, we would map the common fields *first name* to *given name*, *last name* to *surname*, *address* to *address*, and we would ignore the individual fields *phone number*, *occupation*, and *birthdate*.

Once an appropriate mapping has been established, record matching typically proceeds in two stages where (1) individual scores are computed for each pair of corresponding fields across records and (2) individual scores are combined into a single, final match score for the record pair.

3.2.1 Mapped field pairs scoring

The first step in record linkage is to compute similarity scores for each pair of corresponding fields, based on one or a combination of string metrics as discussed above. It is advisable that all returned scores are of the same nature (i.e., either raw or threshold-based) as this simplifies their aggregation in the next stage of record linkage.

If the field matching 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 $[-1,1]$ 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. One of the most common ways to normalize values is by scaling them so that they fall in the interval $[0, 1]$. If the matching score between two values of a field F is x, then x is replaced by $\frac{x-F_{\min}}{F_{\max}-F_{\min}}$, where $F_{\min}$ and $F_{\max}$ are the smallest, respectively, largest, possible values the metric used over F can take.

3.2.2 Score aggregation

Once individual field scores have been computed for all shared fields, they must be combined into a single score representing the overall similarity between the two records of interest. Several aggregation mechanisms are possible, from static ones to more adaptive ones.

Static aggregation techniques rely on a fixed formula. For example, with raw field matching scores, one may take the average value of these scores. If thresholds are used, one may use the Jaccard similarity, defined as the ratio of the number of fields deemed to be matched by their individual field matching algorithm (i.e., their individual scores were above the acceptance threshold) to the total number of common fields (i.e., those involved in the record linkage process). Static approaches to score aggregation are easy to implement, comprehensible and rather efficient. However, they are limited to their fixed form and field comparisons.

Adaptive aggregation techniques generally rely on some form of machine learning [1], wherein information about known matches is used to build predictive models for scoring future record pairs as either matched or unmatched. For example, [36] shows how neural networks can be used to significantly improve record linkage in the context of a large database of genealogical records, while [14] uses neural nets for Sinhala names. Elfeky et al. [6] have also developed a toolkit for record linkage that implements a number of machine learning techniques, including instance-based learning and decision tree induction. The main advantages of adaptive approaches is that they can use features beyond field comparisons and they may bring out predictive patterns that may be missed by humans. On the other hand, these methods require labeled data, which may be costly to obtain; the training of the model is often computationally expensive; and the resulting model may be opaque (e.g., neural networks), making it difficult to understand its decisions.

Finally, we note that, if, as pointed out above, the mapping is correct, but ambiguities remain even after possible standardization, 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. While there may be ways to use machine learning in this context, we leave them as future work and use a static approach here.

3.3 FEC record matching

We now turn to our specific application. The campaign contribution data from which we wish to mine information are distributed across daily reports available through the FEC's FTP site.[1] Our aim is to perform record matching over these daily reports so as to improve the accuracy and validity of statistics derived therefrom. We first describe how individual pairs of donation reports are scored and then how the process can be effectively applied nationwide via a location-based blocking approach.

3.3.1 Scoring donation report pairs

For each individual donation report, the relevant fields or attributes for linkage are name, zip code and street address. Titles and suffixes, such as *Jr* and *Mrs*, as well as all punctuation marks, have been removed from the name field. However, neither names nor street addresses have been standardized in any way. Because of its broad applicability and relative efficiency, we choose to use the Jaro–Winkler string comparison metric for both name and address fields. Since zip codes are in a separate field and linkage takes place across the entire USA, we choose to treat them differently and implement a novel, distance-based metric for zip codes.

The name field is a composite field, and each individual can have a last name, a first name and a middle name. However, because no a priori standardization has been applied, 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 that our data contains two individuals X and Y whose names have been recorded as *Jason S Anderson* and *Anderson Johnson T*, respectively. We begin by building all possible combinations of name components for X and Y and rank them in descending order of their matching scores using the Jaro–Winkler metric, as follows.

[1] See www.fec.gov/finance/disclosure/ftpdet.shtml.

X Component	*Y* Component	Score
Anderson	Anderson	1.00
Jason	Johnson	0.73
Anderson	Johnson	0.69
Jason	Anderson	0.00
Jason	T	0.00
Anderson	T	0.00
S	Anderson	0.00
S	Johnson	0.00
S	T	0.00

Starting from the top of the list, we consider each pair of name components in turn and select it provided that neither of the components it contains has been used in a previous selection. Continuing with X and Y, we would select the first pair (*Anderson*, *Anderson*) and the second pair (*Jason*, *Johnson*). We would then leave out the next six pairs as they each contain at least one component that was part of an earlier selection. Finally, we would select the last pair (S, T). Thus, after alignment, X's name would be rendered as *Anderson Jason S* to be compared against Y's name that would be rendered as *Anderson Johnson T*. Note that the multiset approach assumes that name fields have the same number of components. If one of the name fields has less components than the other, the smaller name field is padded with as many empty strings as are needed to reach the size of the larger field. The multiset approach can then be applied with the assumption that the value of the Jaro–Winkler metric is 0 when either one of its argument is the empty string.

While the above alignment procedure works well in general, there remains a few situations where it fails. Consider again a record containing the name *George Boyd* and another record containing 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 argue that the former is riskier than the latter in our application domain and thus proceed with our multiset approach for names.

Unlike names, addresses are most likely to be expressed in the traditional US format of [number, street name, city, state]. Hence, we choose to avoid the extra computation associated with the multiset approach to re-alignment. The strings making up the addresses of X and Y are compared directly using the Jaro–Winkler metric.

We can now design an aggregate matching score, $\mathrm{sc}(X, Y)$, for X and Y. Let X_N (respectively, Y_N) be the name field for X (respectively, Y) after padding and alignment, and let k be the number of name components in X_N and Y_N. Similarly, let X_A (respectively, Y_A) be the address field for X (respectively, Y). Let $\mathrm{JW}(x, y)$ denote the Jaro–Winkler matching score for strings x and y. Finally, we define $l_N^i = \mathrm{len}(X_N^i) + \mathrm{len}(Y_N^i)$ and $l_A = \mathrm{len}(X_A) + \mathrm{len}(Y_A)$, where $\mathrm{len}(x)$ is the length of string x, i.e., the number of characters in x. Then,

$$\mathrm{sc}(X, Y) = \frac{\sum_{i=1}^{k} l_N^i \mathrm{JW}(X_N^i, Y_N^i) + l_A \mathrm{JW}(X_A, Y_A)}{\sum_{i=1}^{n} l_N^i + l_A} - \mathrm{ZipPenalty}(X, Y)$$

The length-based terms (l_N^i and l_A) act as weights in the numerator and have a normalizing effect in the denominator. The term ZipPenalty(X, Y) is a penalty term based on the map distance between zip code areas. 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.

ZipPenalty(X, Y) is computed as follows. Let ZipX and ZipY be the zip codes of individuals X and Y, respectively. Using a specialized lookup table, we retrieve the longitude and latitude coordinates (ZipX_{lo}, ZipX_{la}) and (ZipY_{lo}, ZipY_{la}) associated with ZipX and ZipY. We then compute the Euclidean distance between them, i.e.,

$$\text{d(ZipX, ZipY)} = \sqrt{(\text{ZipX}_{lo} - \text{ZipY}_{lo})^2 + (\text{ZipX}_{la} - \text{ZipY}_{la})^2}$$

and finally define the penalty term for individuals X and Y as:

$$\text{ZipPenalty}(X, Y) = w.\text{d(ZipX, ZipY)}$$

where w weighs the penalty on the overall matching score. Empirically, $w = 0.002$ was found to give good results.

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 offers a simple mechanism to avoid overlinkage when linking across wide areas such as the entire USA. Although the penalty term is always of some value, it 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). We return to this issue in the next section.

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

$$\text{match}(X, Y) = \begin{cases} 1, & \text{if sc}(X, Y) \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

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

3.3.2 Linking donation reports nationwide

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 10^{14} or 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.

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 [1]. 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 individual X's area overlaps with individual Y's and individual Y's area overlap with individual Z's, we would first link each area independently, but then run into the problem of having to join the results together since there are people in both X's and Y's areas, and people in both Y's and Z's areas. Furthermore, the overlap induced a transitivity problem, similar to the probabilistic overlinkage described above. Indeed, one might easily envisage a situation where two records link between X's and Y's areas, and the record in Y's area also links with one in Z's area, thus creating a link between the individual in X's area and the individual in Z's area. 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 X's area and Z's area. 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 need is a kind of blocking approach, where we can 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 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 USA 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 somewhat 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 4–5 standard PCs and kept cycling MSAs through them as the previous ones would complete. We downloaded the shapefile from the Census's website,[2] which contains all 369 Metropolitan and 578 Micropolitan areas. Zip codes 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

[2] See www.census.gov/geo/www/cob/mmsa2003.html#ascii.

for the 2007–2008 US election cycle are spread over these MSAs with an average size of 9,790 records. The smallest MSA (Guyama, Puerto Rico) contains less than 12 records. The largest MSA (Washington–Arlington–Alexandria) contains in the order of 625,000 records. Instead of the time required to perform 10^{14} comparisons in a sequential implementation, a fully parallel approach requires only the time to perform in the order of 10^{11} comparisons, a dramatic saving (factor of 1,000) in computational time.

4 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.

4.1 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 Web site, which does not contain street addresses. Hence, the computerized linkage is also done here almost exclusively on names. There is a possible implicit address bias induced by the ZipPenalty term, but that bias is small because the records are all localized to individual states. Since address data were not included, no MSA partitioning is used. 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.

- a is the 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 is the 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 is the 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 is the 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 [30]. It may be viewed as a measure of the accuracy of the linkage.

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

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 [16] to compensate for records that may have been linked by chance.

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

Tables 2, 3 and 4 summarize the relationship between manually linked and computer-linked records for New York, Nevada and Utah, respectively.

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 three manually labeled samples are aggregated.

Table 2 Computer versus 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 %)

Table 3 Computer versus 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 versus 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 %)

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

The high values of the Rand index and the adjusted Rand index suggest that there is strong agreement between the computer-generated linkages and 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 Loral Corporation	NY	10021
Schwartz, B. L	New York Loral Space Communications	NY	10021
NELDICH, DAN	NEW YORK GOLDMAN SACHS	NY	10028
Neidich, Dan	New York Goldman Sachs/Managing Partner	NY	10028

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.

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

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

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 RETIRED	NY	10128

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.

4.2 Agreement with self-reported information

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 that 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. The results presented here are based on 1,936 returns from individuals whose name and address information we collected from FEC records. 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. 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 further test the reliability of the matching procedure.

For each major presidential candidate C and for each individual I in our sample, there are four possible outcomes as follows.

- False Negative (FN): There was no record of I's donation to C in our linked database, but I reported giving to C in the survey.
- True Positive (TP): There was a record of I's donation to C in our linked database, and I reported giving to C in the survey.
- False Positive (FP): There was a record of I's donation to C in our linked database, but I did not report giving to C in the survey.
- True Negative (TN): There was no record of I's donation to C in our linked database, and I did not report giving to C in the survey.

For comparison purposes, we focus on rates rather than on raw numbers, namely precision ($P = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$), recall ($R = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$) and false-positive rate ($\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$). Other measures, such as false-negative rate and true-negative rate, are easily derived from these (e.g., $\mathrm{FNR} = 1 - R$, $\mathrm{TNR} = 1 - \mathrm{FPR}$).

Good linkage would exhibit high precision, high recall and low false precision rate. A couple of remarks about FN and FP are in order in this respect before analyzing our results, since comparisons between observed behavior (from the linkage) and self-reported behavior (from the survey) are complicated somewhat by the FEC reporting requirements and bias in our survey process.

- *FEC reporting rule*. Because contributions are typically not disclosed until they reach the $200 threshold, it is difficult to 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 three $75 donations to Obama and who was not disclosed for two $75 donations to Biden, yet reported giving to both candidates in their survey, would come up as a true positive for Obama and a false negative for Biden. While some committees voluntarily disclosed smaller donors, they represented a very small proportion in our data. Individual donors, on the other hand, are likely to simply list all of the candidates to whom they contributed irrespective of the amount given. As a result, the values reported here for FN are likely overstated, thus negatively impacting recall.
- *Survey bias*. Individuals were asked to indicate only which candidates they contributed to. As a result, when a candidate's name is not indicated on a survey, we cannot be sure that the individual did not donate to that candidate, only that they did not report giving to that candidate. There is indeed a subtle but important distinction between reporting not giving to a candidate and not reporting giving to a candidate. Our survey measures only

Table 6 Linkage versus self-reports for major presidential candidates

	P	*R*	F PR
Biden	0.83	0.15	0.001
Clinton	0.86	0.49	0.006
Edwards	0.82	0.34	0.005
Giuliani	0.84	0.36	0.005
Huckabee	1.00	0.39	0.000
McCain	0.99	0.73	0.007
Obama	0.99	0.74	0.008
Paul	0.94	0.76	0.003
Richardson	0.83	0.29	0.002
Romney	0.92	0.40	0.005

the former and assumes that the latter is the same as the former. As a result, the values recorded for FP are likely inflated, thus impacting precision and false-positive rate, and suggesting a larger error in linkage than is the case. As a matter of fact, a closer examination of our data reveals that the greatest portion of the false positives arises from individuals in the sample who claimed not to have contributed to any candidates. This may be due to concerns over privacy (i.e., unwillingness to disclose information about contributions in the survey), reluctance to admit giving to losing candidates, mere omissions or other reasons.[3]

While there is no way to remove the survey bias, we can improve the assessment of the linkage process by excluding from the survey all of the individuals who reported not giving any contributions. Of the 1,936 individuals considered here, there are 293 such individuals, leaving a final sample of 1,643 individuals. Table 6 shows the values of precision, recall and false-positive rate for the major presidential candidates for these individuals.[4]

These results are strong evidence for the validity of our linkage process. Precision is high (i.e., above 82 %) for all candidates and false precision rate is low, at less than 0.5 % for all candidates. The results are particularly strong for the two main candidates (i.e., Obama and McCain) since they arise from larger pools of donors (TP = 465 for McCain and TP = 588 for Obama).

5 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. 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." [32]. 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

[3] The privacy concern may actually be quite prevalent as these same individuals (found in our linkage but who did not report giving to any candidates in the survey) are also more than twice as likely as others not to report their income.

[4] Even when the "offending" individuals are not removed, FPR does not exceed 0.039 and precision does not go below 0.71 for any of the candidates.

Table 7 Donation/donor summary statistics (in dollars) for 2007–2008

	Mean	Median	SD
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

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 come from two complementary sources, as follows.

1. *FEC Records*. As mentioned above, 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. Note that 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.
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). 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. We had to sign appropriate non-disclosure agreements to obtain these data.

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.

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

Table 8 Percentage distribution of Obama and McCain's donations and donors by amount

Amount	Unlinked		Linked	
	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

Table 9 Percentage distribution of Obama and McCain's total itemized individual donations raised for donations and donors of different sizes

Amount	Unlinked		Linked	
	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

specific campaigns, with 48 % and 35 % higher values for Obama and McCain, respectively. Furthermore, the candidate-specific data suggest that Obama seems to have attracted more repeat small donors than McCain did. We take a closer look at the differences between 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.

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.

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 Figs. 1, 2, 3 and 4 provide another view of these effects (Figs. 1, 2 are based on the CSR data). They are histograms of the distributions of donations versus 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 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 column (n) reports the number of donors (donations). The second column (%O) and third column (%M) 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

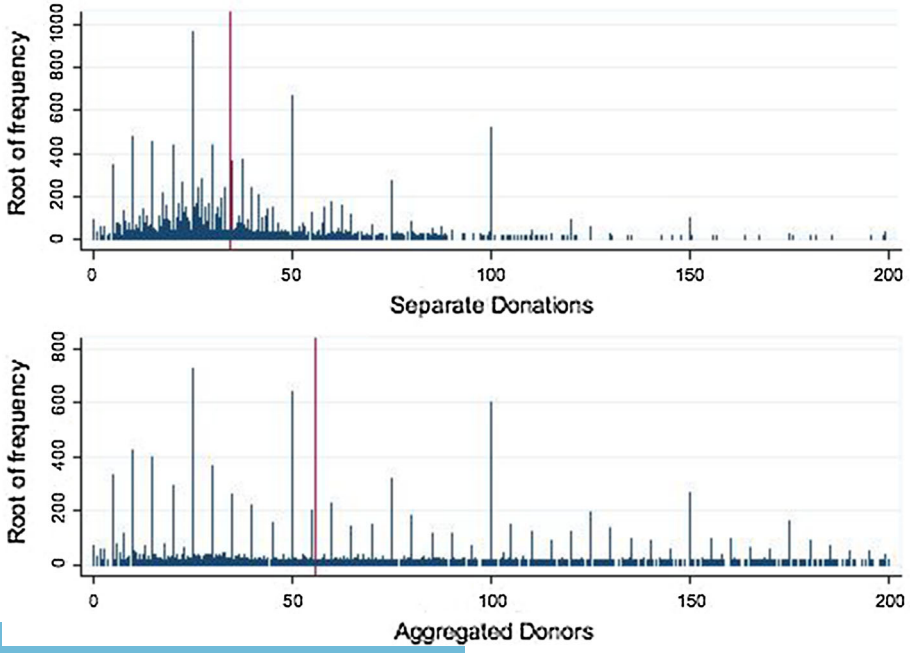

Fig. 1 Small donations versus donors for Obama

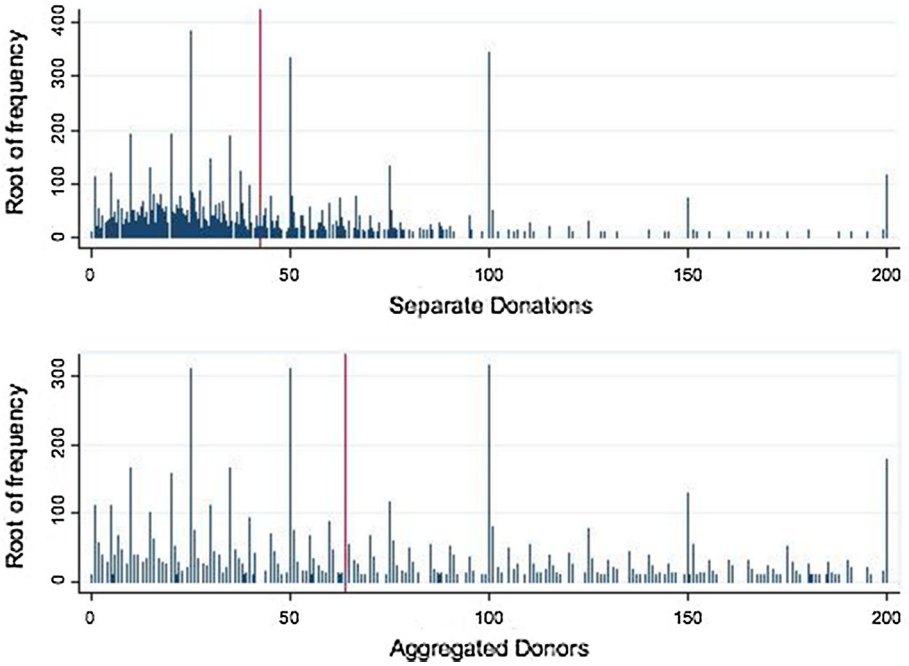

Fig. 2 Small donations versus donors for McCain

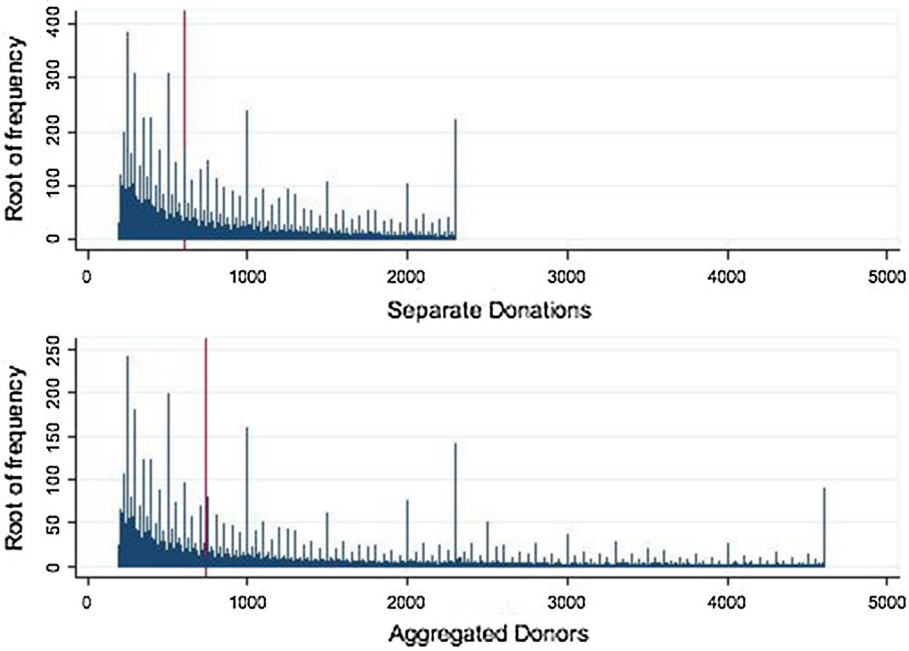

Fig. 3 Large donations versus donors for Obama

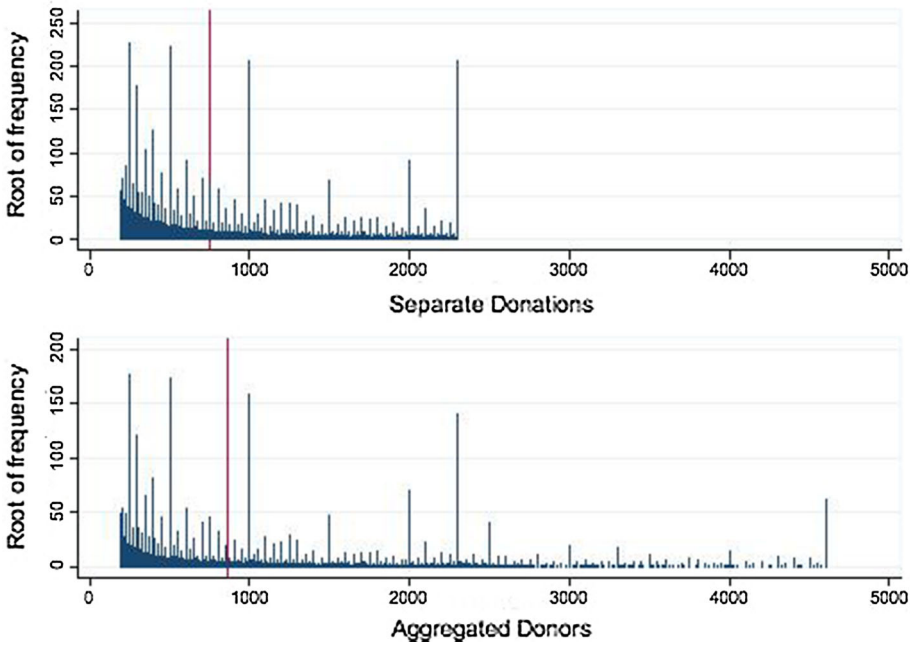

Fig. 4 Large donations versus donors for McCain

Table 10 Description of contributors (Contributions) by amount given

	n	%O	%M	#D	%H	%C
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)

McCain, and 5 % of all donors who gave in that range donated to McCain. The fourth column (#D) counts the mean number of distinct contributions made by donors. The fifth column (%H) examines what percentage of donors (donations) contributed to House candidates, as opposed to Senate, presidential, PACs, and parties. Finally, the last column (%C) 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 when mining donations, donors and campaign results. The following are a few observations based on Table 10.

- 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. 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.
- While large donations are rarely given to House candidates, large donors 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 substantiate claims made when mining data about campaign activities and results, when donors are to be taken as the unit of analysis.

6 Linking AidData projects

To further illustrate the value of record linkage in political studies, we briefly discuss how this technique may be used in the context of AidData projects.[5] AidData publishes information about development assistance finance, including detailed data about specific donation projects over many years. In order to prepare datasets to be published, information from different sources about these donation projects must be adequately linked.

Each individual project in AidData is assigned a unique project number. When several instances of the same project share the same project number, then linkage is trivial. However, project numbers are not always accurate, which makes it difficult to get a clear picture of

[5] See www.aiddata.org/content/index.

certain projects. For this illustration, the dataset we focus on is one that was input manually. Because research assistants typed in the data by hand and the original documents are sometimes difficult to read, one can expect data entry errors. We use record linkage to identify project numbers that were probably incorrect.

The data were made available in Excel format for the years 1965–1974 and 1980–1991. Each row had information about a particular project, described by 27 fields including Project Number, Project Title, Account Name, and Recipient, as well as financial data such as Total Project Planned Cost, Estimated Obligations and Actual Expenditures. Multiple rows made reference to the same project at different points in time and were thus the target of record linkage. We elected to do the linkage using only a subset of the fields as not all fields were consistent across rows. For example, financial data tend to change from year to year. The fields that were deemed the most useful in the linking process, and were thus considered by our algorithm, were Recipient, Project Title, Project Number and FY Initial Obligation.

If two records belong to the same project, then they would most likely have very similar values for these four fields. For example, the two records shown below have the same recipients, titles and initial obligation years; however, the project numbers are different. These two records should have the same project number; an error was likely made during data entry.

Project number	Recipient	Project title	Obligation year
5150175	Costa Rica	Coop Banking Services and Credit	1983
5150178	Costa Rica	Coop Banking Services and Credit	1983

In order to detect such errors, we look for pairs of records that appear very similar, but have different project numbers. Records that look the same probably belong to the same project; therefore, they probably should have matching project numbers. If two records look similar, yet have different project numbers, then we flag the records as errors. We find pairs of similar records using the field matching and record matching methods used for the FEC linking.

The matching is accomplished by comparing each record with every other record in the dataset. When comparing two records, we first compare the recipient fields of the records. If the recipients do not match exactly, then we discard the pair and do not count it as a match. Otherwise, we continue to compare the records. We then give the pair a score based on how similar the project title, project number and initial obligation year fields of the two records are. The first step in computing this score is to find the Jaro–Winkler matching scores for each of the three fields. Then, the final score between records X and Y is computed as follows.

$$\mathrm{sc}(X, Y) = \frac{\mathrm{JW}(X_{\mathrm{title}}, Y_{\mathrm{title}}) + \mathrm{JW}(X_{\mathrm{number}}, Y_{\mathrm{number}}) + \mathrm{JW}(X_{\mathrm{year}}, Y_{year})}{3}$$

where, as above, $\mathrm{JW}(x, y)$ is the Jaro–Winkler matching score of the strings x and y. As in the FEC linking, we threshold this raw score to determine whether or not a pair is a match. Here, the acceptance threshold is set at 0.9. If a pair's score is above 0.9, then we determine that the pair is a match. Once we determine which records likely belong to the same project, we compare their project numbers. If we believe that two records belong to the same project, but the records have different project numbers, then we add the pair to a list of possible errors. These records can then be compared to the original documents to determine whether a data entry error has occurred.

This error-detection process results in pairs of records, but the error is probably only in one of the records. Hence, after finding a flagged pair of records, we calculate which of the records is most likely to contain the error. This is done by comparing each of the records in the pair with

the other records that shared its project number. If a record is very similar to the other records with the same project number, then it is probably not an error. If, on the other hand, the record is different from the other records with the same project number, then it most likely has the wrong project number. We determine similarity between a record and other records with its same project number using the Jaro–Winkler matching score. In each flagged pair, the record that is least similar to other records with its project number is labeled as the error. As with FEC linking, this semi-automated linkage process greatly simplifies the work of researchers and increases the quality and reliability of conclusions drawn from analyses of the data.

7 Discussion and conclusion

The method presented here builds on previous work in record linkage and shows how record linkage techniques can be adapted to the domain of political science, especially in the challenging situation where limited information is available (here, mainly name and address), and data must be linked across a very wide geographical area (here, the entire USA). The success of our approach is based on the exploitation of domain knowledge together with technical innovations in the linkage metric and process, as follows.

- *Multiset Distance for Names*. One of the challenges in dealing with names in record linkage is the accurate classification of first, middle and last names, especially when a number of names are ambivalent and can thus be both a first name and a last name (e.g., Lloyd, Spencer, Morgan). A typical approach in record linkage is to use standardization to try to recover and align the pieces of names within a record. This process is expensive and error-prone. Instead, we use a simple multiset approach wherein we compare every name piece in one record with every name piece in the other and retain the highest scoring matching pieces. While this approach is still not 100% accurate, it alleviates most issues in an automatic and efficient way.
- *Euclidean Distance for Zip Codes*. In most applications, zip codes are an integral part of the address field and treated in aggregate with the rest of the address. Even when they are considered separately, they are generally still treated as a string. In both cases, zip codes are thus compared by standard string metrics. While this may be efficient, it cannot exploit the inherent geographical information available in zip codes, such as how close or far apart two areas may be. Because there is so much redundancy in street addresses across the USA (e.g., many cities use a standard grid system, many historical figures are used to name streets), there is significant information in zip codes that may be lost when they are considered as strings. The first two digits are indeed unique to each state, but there is no further geographical correspondence beyond them. For example, the cities of Orem and Provo in Utah border on each other, yet their zip codes are very different: Orem's zip codes are 84057, 84058, 84059 and 84097, while Provo zip codes are 84601, 84602, 84603, 84604, 84604, and 84605. Conversely, the city of Madison, Ohio, is located 1,783 miles from Orem, yet its zip code of 44057 differs in only one digit from that of Orem. As a result, we treat zip codes as geographical locations, use physical distances as their level of similarity and penalize record matching scores based on these distances. Hence, if two records tend to match string-wise, but they are geographically far apart as indicated by their zip codes, their matching score is reduced. This mechanism reduces the risk of over-linkage.
- *Domain Awareness and MSA*. The linking of political data, such as campaign donations, is inherently about individuals. Hence, we factor human behavior in our linkage process. We recognize that individuals may be in different places at different times (e.g., work vs.

home), yet, that for obvious reasons, such as time to commute, these places are most likely to remain relatively close geographically. This aspect of human behavior has been captured by the United States Office of Management and Budget through MSAs. Hence, we propose the use of MSAs to "bound" the geography within which an individual is expected to be found, and essentially to serve as an efficient blocking mechanism in our record linkage. There are two principal advantages to using MSAs that contribute to better results:

- Because MSAs are independent from each other (i.e., there is no geographical overlap), effecting linkage across the entire USA becomes feasible through parallel computation.
- Because MSAs capture natural human behavior, they avoid undesirable transitivity and thus, together with the zip penalty, significantly reduce the risk of over-linkage.

We applied the proposed 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. We also showed how the same idea of record linkage may be used to assist researchers in linking other disparate sources of political data, as a preliminary step to mining the aggregated data more accurately.

The approach described here should prove useful in other application domains such as linking individuals and families across US censuses, for example to study patterns of population migration. On the technical side, it is possible that results may be further improved by first optimizing the metric field for each field. Furthermore, since we have labeled data available (based on manual linkage), it would be interesting to try to use machine learning techniques to link the FEC data.

Acknowledgments Our thanks to Yao Huang, Weston Rowley, David Wilcox and David Lassen for research assistance and computer code. We are also grateful to David Magleby and Joseph Olson for their support, encouragement, and advice. Finally, we thank the anonymous reviewers for their very useful comments and suggestions.

References

1. Bilenko M, Mooney R, Cohen W, Ravikumar P, Fienberg S (2003) Adaptive name matching in information integration. IEEE Intel Syst 18(5):16–23
2. Cheatham M, Hitzler P (2013) String similarity metrics for ontology alignment. In: Proceedings of the twelfth international semantic Web conference (LNCS 8219), pp 294–309
3. Christen P (2006) A comparison of personal name matching: techniques and practical issues. Technical Report TR-CS-06-2, Department of Computer Science, The Australian National University
4. Christen P (2012) Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer, Berlin
5. Cohen W, Ravikumar P, Fiendberg S (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
6. Elfeky MG, Verykios VS, Elmagarmid AK, Ghanem TM, Huwait AR (2003) Record linkage: a machine learning approach, a toolbox, and a digital government Web service. Technical Report 03–024, Department of Computer Science, Purdue University
7. Elmagarmid A, Ipeitoris P, Verykios V (2007) Duplicate record detection: a survey. IEEE Trans Knowl Data Eng 19(1):1–16
8. Fellegi I, Sunter A (1969) A theory for record linkage. J Am Stat Assoc 64(328):1183–1210

9. Fu Z, Christen P, Boot M (2011) Automatic cleaning and linking of historical census data using household information. In: Proceedings of the IEEE eleventh international conference on data mining workshops, pp 413–420
10. Fu Z, Christen P, Zhou J (2014) A Graph Matching Method for Historical Census Household Linkage. In: Proceedings of the eighteenth Pacific-Asia conference on knowledge discovery and data mining (LNAI 8443), pp 485–496
11. Gadd T (1990) PHONIX : the algorithm. Prog Autom Library Inform Syst 24(4):363–366
12. Gu L, Baxter R, Vickers D, Rainsford C (2003) Record linkage: current practice and future directions. Tech. Rep. No. 03/83, CSIRO Mathematical and Information Sciences
13. Herzog TH, Scheuren F, Winkler WE (2010) Record Linkage. Wiley Interdiscip Rev Comput Stat 2(5):535–543
14. Hettiarachchi GP, Attygalle D, Hettiarachchi DS, Ebisuya A (2013) A generic statistical machine learning and data mining framework for record classification and linkage. Int J Intel Inform Process 4(2):96–106
15. Howe GR, Lindsay J (1981) A generalized iterative record linkage computer system for use in medical follow-up studies. Comput Biomed Res 14(4):327–340
16. Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
17. Irvine KA, Taylor LK (2011) The Centre for Health Record Linkage: fostering population health research in NSW. NSW Pub Health Bull 22(2):17–18
18. Ivie S, Pixton B, Giraud-Carrier C (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
19. Jaro M (1995) Probabilistic linkage of large public health data file. Stat Med 14(5–7):491–498
20. Lain SJ, Algert CS, Tasevski V, Morris JM, Roberts CL (2009) Record linkage to obtain birth outcomes for the evaluation of screening biomarkers in pregnancy: a feasibility study. BMC Med Res Methodol 9:48
21. Lait A, Randell B (1993) An assessment of name matching algorithms. Department of Computer Science, University of Newcastle upon Tyne, UK, Tech. rep
22. Levenshtein V (1966) Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Doklady 10:707–710
23. Monge A, Elkan C (1996) The field-matching problem: algorithm and applications. In: Proceedings of the second international conference on knowledge discovery and data mining, pp 267–270
24. Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
25. Newcombe H, Kennedy J, Axford S, James A (1959) Automatic linkage of vital records. Science 130(3381):954–959
26. Pfeifer U, Poersch T, Fuhr N (1996) Retrieval effectiveness of proper name search methods. Inf Process Manag 32(6):667–679
27. Philips L (2000) The double-metaphone search algorithm. C/C++ Users J 18(6):38–43
28. Pixton B, Giraud-Carrier C (2005) MAL4:6 - Using data mining for record linkage. In: Proceedings of the 5th annual Workshop on technology for family history and genealogical research
29. Quass D, Starkey P (2003) Record Linkage for Genealogical Databases. In: Proceedings of the ACM SIGKDD workshop on data cleaning, record linkage, and object consolidation
30. Rand W (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
31. Ruggles S (2002) Linking historical censuses: a new approach. Hist Comput 14(1+2):213–224
32. Solomon J (2007) Obama takes lead in money raised. Washington Post, July 2:A1
33. Stavrou EP, Baker DF, Bishop JF (2009) Maternal smoking during pregnancy and childhood cancer in New South Wales: a record linkage investigation. Cancer Causes Control 20(9):1551–1558
34. St. Sauver JL, Grossardt BR, Yawn BP, Melton LJ 3rd, Pankratz JJ, Brue SM, Rocca WA (2012) Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. Int J Epidemiol 41(6):1614–1624
35. Sweet C, Odyer T, Alhajj R (2007) Enhanced graph based genealogical record linkage. In: Proceedings of the third international conference on advanced data mining and applications (LNAI 4632), pp 476–487
36. Wilson DR (2011) Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In: Proceedings of the international joint conference on neural networks, pp 9–14
37. Winkler WE (2001) Record linkage software and methods for merging administrative lists. Statistical research report series No. RR2001/03. http://www.vrdc.cornell.edu/info7470/2011/Readings/rr2001-03
38. Winkler W (2006) Overview of record linkage and current research directions. Research Report Series (Statistics #2006-2). http://www.census.gov/srd/papers/pdf/rrs2006-02
39. Zobel J, Dart P (1995) Finding approximate matches in large lexicons. Softw Pract Exp 1:331–345

C. Giraud-Carrier is Associate Professor and coordinator of the Data Mining laboratory in the Department of Computer Science at Brigham Young University (BYU). Prior to joining BYU, he was Senior Manager at ELCA, a Swiss IT services company, where his responsibilities included the capitalization of Data Mining expertise. Prior to this, he was Senior Lecturer in the Department of Computer Science at the University of Bristol, where he founded and led the Machine Learning Research Group. Dr Giraud-Carrier holds a Ph.D. in Computer Science from BYU.

J. Goodliffe is an Associate Professor of Political Science, and Research Scholar at the Center for the Study of Elections and Democracy, at Brigham Young University. He holds a Ph.D. in political science from the University of Rochester and a SB in Aeronautics and Astronautics from MIT. Substantively, his research interests include international human rights treaties, democratization, and US campaigns and elections emphasizing campaign finance. Methodologically, his interests include record linkage, social network analysis, game theory, and multilevel/hierarchical models.

B. M. Jones is a Ph.D. candidate at the University of Wisconsin studying political science. He holds a M.A. in political science from the University of Wisconsin and a B.A. in political science from Brigham Young University. His research interests include political psychology, political participation, computational social science and methods for the quantitative analysis of texts.

S. Cueva holds a B.S. in Computer Science from Brigham Young University. Her research in the BYU Data Mining Lab included record linkage, community mining and social network analysis. She is currently working in the software industry as a software engineer.