Record Linking Examples

John M. Abowd and Lars Vilhuber
March 2005
Need for *automated* record linkage

- RA time required for the following matching tasks:
  - Finding financial records for Fortune 100: 200 hours
  - Finding financial records for 50,000 small businesses: ?? hours
  - Unduplication of the U.S. Census survey frame (115,904,641 households): ???
  - Identifying miscoded SSNs on 500 million wage records: ???
  - Longitudinally linking the 12 million establishments in the Business Register: ???
Basic methodology

- Name and address parsing and standardization
- Identifying comparison strategies:
  - Which variables to compare
  - String comparator metrics
  - Number comparison algorithms
  - Search and blocking strategies
- Ensuring computational feasibility of the task
Generic workflow

• Standardize
• Match
• Revise and iterate through again
An example


• Approx. 500 million records (quarterly wage records for 1991-1999, California)
• 28 million SSNs
SSN Name editing

Example

Coded Name | Coded SSN | EIN | Earnings
---|---|---|---
Leslie Kay | 1 | A | $10
Leslie Kay | 21 | A | $10
Lesly Kai | 31 | B | $11

1’s tenure with A:
1’s employment history

Separations too high
Accessions too high

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
Need for Standardization

• Names may be written many different ways
• Addresses can be coded in many different ways
• Firm names can be formal, informal, or differ according to the reporting requirement
How to standardize

• Inspect the file to refine strategy
• Use commercial software
• Write custom software (SAS, Fortran, C)
• Apply standardizer
• Inspect the file to refine strategy
Standardizing Names

• Alternate spellings

1. Dr. William J. Smith, MD
2. Bill Smith
3. W. John Smith, MD
4. W.J. Smith, Jr.
5. Walter Jacob Smith, Sr.
# Standardized names

<table>
<thead>
<tr>
<th>Pre</th>
<th>First</th>
<th>Mid</th>
<th>Last</th>
<th>Pos 1</th>
<th>Post 2</th>
<th>Alt1</th>
<th>Std1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dr</td>
<td>William</td>
<td>J</td>
<td>Smith</td>
<td>MD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bill</td>
<td>Smith</td>
<td></td>
<td></td>
<td></td>
<td>William</td>
<td>BWILL</td>
</tr>
<tr>
<td>3</td>
<td>W</td>
<td>John</td>
<td>Smith</td>
<td>MD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>W</td>
<td>J</td>
<td>Smith</td>
<td>Jr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Walter</td>
<td>Jacob</td>
<td>Smith</td>
<td>Sr</td>
<td></td>
<td></td>
<td>WALT</td>
</tr>
</tbody>
</table>
Standardizing addresses

- Many different pieces of information
  1. 16 W Main Street #16
  2. RR 2 Box 215
  3. Fuller Building, Suite 405, 2nd door to the right
  4. 14588 Highway 16W
### Standardized addresses

<table>
<thead>
<tr>
<th>Pre</th>
<th>Hsnp</th>
<th>Stnm</th>
<th>RR</th>
<th>Box</th>
<th>Post1</th>
<th>Post2</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Bldg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>W</td>
<td>16</td>
<td>Main</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td>215</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
<td>215</td>
<td></td>
<td></td>
<td></td>
<td>405</td>
<td>Fuller</td>
</tr>
<tr>
<td>4</td>
<td>14588</td>
<td>Hwy</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A&V: standardizing

• Knowledge of structure of the file: -> No standardizing
• Matching will be within records close in time -> assumed to be similar, no need for standardization
• BUT: possible false positives -> chose to do an weighted unduplication step (UNDUP) to eliminate wrongly associated SSNs
A UID is a unique combination of SSN-First-Middle-Last

<table>
<thead>
<tr>
<th>SSN</th>
<th>UID</th>
<th>First</th>
<th>Middle</th>
<th>Last</th>
<th>Earn</th>
<th>YQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>123-45-6789</td>
<td>58</td>
<td>John</td>
<td>C</td>
<td>Doe</td>
<td>25678</td>
<td>93Q1</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>58</td>
<td>John</td>
<td>C</td>
<td>Doe</td>
<td>26845</td>
<td>93Q2</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>59</td>
<td>Jon</td>
<td>C</td>
<td>Doe</td>
<td>24837</td>
<td>94Q4</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>60</td>
<td>Robert</td>
<td>E</td>
<td>Lee</td>
<td>7439</td>
<td>93Q1</td>
</tr>
<tr>
<td>SSN</td>
<td>UID</td>
<td>First</td>
<td>Middle</td>
<td>Last</td>
<td>Earn</td>
<td>YQ</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>58</td>
<td>John</td>
<td>C</td>
<td>Doe</td>
<td>25678</td>
<td>93Q1</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>58</td>
<td>John</td>
<td>C</td>
<td>Doe</td>
<td>26845</td>
<td>93Q2</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>59</td>
<td>Jon</td>
<td>C</td>
<td>Doe</td>
<td>24837</td>
<td>94Q4</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>60</td>
<td>Robert</td>
<td>E</td>
<td>Lee</td>
<td>7439</td>
<td>93Q4</td>
</tr>
<tr>
<td>123-45-6789</td>
<td>60</td>
<td>Robert</td>
<td>E</td>
<td>Lee</td>
<td>7439</td>
<td>94Q1</td>
</tr>
</tbody>
</table>

Conservative strategy: Err on the side of caution
## A&V: UNDUP (3)

<table>
<thead>
<tr>
<th>SSN</th>
<th>UID</th>
<th>First</th>
<th>Middle</th>
<th>Last</th>
<th>Earn</th>
<th>YQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>123-54-6789</td>
<td>38</td>
<td>Roberta</td>
<td>C</td>
<td>Doe</td>
<td>25678</td>
<td>93Q1</td>
</tr>
<tr>
<td>123-54-6789</td>
<td>38</td>
<td>Roberta</td>
<td>C</td>
<td>Doe</td>
<td>26845</td>
<td>93Q2</td>
</tr>
<tr>
<td>123-54-6789</td>
<td>39</td>
<td>Roberta</td>
<td></td>
<td>Doe</td>
<td>24837</td>
<td>94Q4</td>
</tr>
<tr>
<td>123-54-6789</td>
<td>40</td>
<td>Bobbie</td>
<td></td>
<td>Lee</td>
<td>27439</td>
<td>93Q4</td>
</tr>
<tr>
<td>123-54-6789</td>
<td>40</td>
<td>Bobbie</td>
<td></td>
<td>Lee</td>
<td>27439</td>
<td>94Q1</td>
</tr>
</tbody>
</table>

Conservative strategy: Err on the side of caution

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
Matching

- Define match blocks
- Define matching parameters: marginal probabilities
- Define upper $T_u$ and lower $T_l$ cutoff values
Record Blocking

• Computationally inefficient to compare all possible record pairs
• Solution: Bring together only record pairs that are LIKELY to match, based on chosen blocking criterion
• Analogy: SAS merge by-variables
Blocking example

• Without blocking: $A \times B$ is $1000 \times 1000 = 1,000,000$ pairs

• With blocking, f.i. on 3-digit ZIP code or first character of last name. Suppose 100 blocks of 10 characters each. Then only $100 \times (10 \times 10) = 10,000$ pairs need to be compared.
A&V: Blocking and stages

• Two stages were chosen:
  – UNDUP stage (preparation)
  – MATCH stage (actual matching)

• Each stage has own
  – Blocking
  – Match variables
  – Parameters
A&V: UNDUP blocking

• No comparisons are ever going to be made outside of the SSN
• Information about frequency of names may be useful
• Large amount of records: 57 million UIDs associated with 28 million SSNs, but many SSNs have a unique UID

⇒ Blocking on SSN
⇒ Separation of files by last two digits of SSN (efficiency)
A&V: MATCH blocking

• Idea is to fit 1-quarter records into work histories with a 1-quarter interruption at same employer

⇒ Block on Employer – Quarter

⇒ Possibly block on Earnings deciles
A&V: MATCH block setup

# Pass 1:
BLOCK1 CHAR SEIN SEIN
BLOCK1 CHAR QUARTER QUARTER
BLOCK1 CHAR WAGEQANT WAGEQANT
# follow 3 other BLOCK passes with identical setup
#
# Pass 2: relax the restriction on WAGEQANT
BLOCK5 CHAR SEIN SEIN
BLOCK5 CHAR QUARTER QUARTER
# follow 3 other BLOCK passes with identical setup
Determination of match variables

• Must contain relevant information
• Must be informative (distinguishing power!)
• May not be on original file, but can be constructed (frequency, history information)
A&V: Variables and Matching

• File only contains Name, SSN, Earnings, Employer
• Construct frequency of use of name, work history, earnings deciles
• Stage 1: use name and frequency
• Stage 2: use name, earnings decile, work history with employer
Understanding comparators

• Comparators need to account for
  – Typographical error
  – Significance of slight variations in numbers (both absolute and relative)
  – Possible variable inversions (first and last name flipped)
String comparators: Soundex

• The first letter is copied unchanged
• Subsequent letters:
  - bfpv -> "1"
  - cgjkqszxç -> "2"
  - dt -> "3"
  - l -> "4"
  - mnñ -> "5"
  - r -> "6"
• Other characters are ignored
• Repeated characters treated as single character.
• 4 chars, zero padded.
For example, "SMITH" or "SMYTHE" would both be encoded as "S530".

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
String comparators: Jaro

• First returns a value based on counting insertions, deletions, transpositions, and string length
• Total agreement weight is adjusted downward towards the total disagreement weight by some factor based on the value
• Custom adjustments (Winkler and others)
Comparing numbers

• A difference of “34” may mean different things:
  – Age: a lot (mother-daughter? Different person)
  – Income: little
  – SSN or EIN: no meaning

• Some numbers may be better compared using string comparators
Number of matching variables

• In general, the distinguishing power of a comparison increases with the number of matching variable

• Exception: variables are strongly correlated, but poor indicators of a match

• Example: General business name and legal name associated with a license.
Determination of match parameters

• Need to determine the conditional probabilities $P(\text{agree}|M)$, $P(\text{agree}|U)$ for each variable comparison

• Methods:
  – Clerical review
  – Straight computation (Fellegi and Sunter)
  – EM algorithm (Dempster, Laird, Rubin, 1977)
  – Educated guess/experience
  – For $P(\text{agree}|U)$ and large samples (population): computed from random matching
Determination of match parameters (2)

- Fellegi & Sunter provide a solution when $\gamma$ represents three variables. The solution can be expressed as marginal probabilities $m_k$ and $u_k$
- In practice, this method is used in many software applications
- For $k>3$, method-of-moments or EM methods can be used.
Marginal probabilities: educated guesses for starting values

- $P(\text{agree on characteristic } X | M) =$
  - 0.9 if $X =$ first, last name, age
  - 0.8 if $X =$ house no., street name, other characteristic

- $P(\text{agree on characteristic } X | U) =$
  - 0.1 if $X =$ first, last name, age
  - 0.2 if $X =$ house no., street name, other characteristic

Note that distinguishing power of first name ($R(\text{first}) = 0.9/0.1 = 9$) is larger than the street name ($R(\text{street}) = 0.8/0.2 = 4$)
Marginal probabilities: better estimates of $P(\text{agree}|M)$

• $P(\text{agree} | M)$ can be improved after a first match pass by a clerical review of match pairs:
  – Draw a sample of pairs
  – Manual review to determine “true” match status
  – Recompute $P(\text{agree}|M)$ based on known truth sample
A&V: UNDUP match variables

# Pass1
MATCH1 NAME_UNCERT namef 0.9 0.001 700
MATCH1 NAME_UNCERT namel 0.9 0.02 700
MATCH1 NAME_UNCERT namem 0.9 0.02 700
MATCH1 NAME_UNCERT concat 0.9 0.02 700

# Pass 2
MATCH2 ARRAY NAME_UNCERT fm_name 0.9 -.02 750
MATCH2 NAME_UNCERT namel 0.9 0.001 700
MATCH2 NAME_UNCERT concat 0.9 0.02 700
# and so on...
A&V: MATCH match variables

# Pass1
MATCH1 CNT_DIFF SSN SSN 0.9 0.000001 5
MATCH1 NAME_UNCERT namef namef 0.9 0.02 700
MATCH1 NAME_UNCERT namel namem 0.9 0.02 700
MATCH1 NAME_UNCERT namel namel 0.9 0.001 700  
# Pass 2
MATCH2 CNT_DIFF SSN SSN 0.9 0.000001 5
MATCH2 NAME_UNCERT concat concat 0.9 0.02 700 
# Pass 3
MATCH3 UNCERT SSN SSN 0.9 0.000001 700 
MATCH3 NAME_UNCERT namef namef 0.9 0.02 700
MATCH3 NAME_UNCERT namem namem 0.9 0.02 700 
MATCH3 NAME_UNCERT namel namel 0.9 0.001 700 and so on...

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
Adjusting $P(\text{agree}|M)$ for relative frequency

- Further adjustment can be made by adjusting for relative frequency (idea goes back to Newcombe (1959) and F&S (1969))
  - Agreement of last name by Smith counts for less than agreement by Vilhuber
- Default option for some software packages
- Requires assumption of strong assumption about independence between agreement on specific value states on one field and agreement on other fields.
A&V: Frequency adjustment

• UNDUP:
  – none specified

• MATCH:
  – allow for name info,
  – disallow for wage quantiles, SSN
Marginal probabilities: better estimates of $P(\text{agree} | U)$

- $P(\text{agree} | U)$ can be improved by computing random agreement weights between files $\alpha(A)$ and $\beta(B)$ (i.e. $A \times B$)
  - # pairs agreeing randomly by variable $X$ divided by total number of pairs
Error rate estimation methods

- Sampling and clerical review
  - Within L: random sample with follow-up
  - Within C: since manually processed, “truth” is always known
  - Within N: Draw random sample with follow-up. Problem: sparse occurrence of true matches
  - Model the shape of the matching weight distributions (empirical density of R) if sufficiently separated
- Capture-recapture with different blocking for false non-match rates
Analyst Review

• Matcher outputs file of matched pairs in decreasing weight order
• Examine list to determine cutoff weights and non-matches.
A&V: Finding cutoff values

- **UNDUP:**
  - CUTOFF1 7.5 7.5
  - CUTOFF2 8 8
  - Etc.

- **MATCH:**
  - CUTOFF1 18 18
  - CUTOFF2 12 12
  - CUTOFF 10 10
  - Etc.
# A&V: Sample matcher output

<table>
<thead>
<tr>
<th>RESULT</th>
<th>RECNUM</th>
<th>WGT</th>
<th>SSN</th>
<th>NAMEF</th>
<th>NAMEM</th>
<th>NAMEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>[UA]</td>
<td>504</td>
<td>-999.99</td>
<td>382661272</td>
<td>WILL</td>
<td>TARY</td>
<td></td>
</tr>
<tr>
<td>[UB]</td>
<td>2827</td>
<td>-999.99</td>
<td>384883394</td>
<td>RICHARD</td>
<td>PHOUK</td>
<td></td>
</tr>
<tr>
<td>[UB]</td>
<td>392</td>
<td>-999.99</td>
<td>335707385</td>
<td>MONA</td>
<td>LISA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESULT</th>
<th>RECNUM</th>
<th>WGT</th>
<th>SSN</th>
<th>NAMEF</th>
<th>NAMEM</th>
<th>NAMEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>[CA]</td>
<td>351</td>
<td>3.66</td>
<td>333343734</td>
<td>DONNA L</td>
<td>DUK</td>
<td></td>
</tr>
<tr>
<td>[CB]</td>
<td>1551</td>
<td>3.66</td>
<td>333383832</td>
<td>MARGEN L</td>
<td>PRODUCT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESULT</th>
<th>RECNUM</th>
<th>WGT</th>
<th>SSN</th>
<th>NAMEF</th>
<th>NAMEM</th>
<th>NAMEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>[MA]</td>
<td>43</td>
<td>32.76</td>
<td>4444444441</td>
<td>LUKE</td>
<td>UPP</td>
<td></td>
</tr>
<tr>
<td>[MB]</td>
<td>169</td>
<td>32.76</td>
<td>4444444447</td>
<td>LUKE</td>
<td>UPP</td>
<td></td>
</tr>
</tbody>
</table>

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
Post-processing

• Once matching software has identified matches, further processing may be needed:
  – Clean up
  – Carrying forward matching information
  – Reports on match rates
Generic workflow (2)

• Start with initial set of parameter values
• Run matching programs
• Review moderate sample of match results
• Modify parameter values (typically only $m_k$) via ad hoc means
Acknowledgements

• This lecture is based in part on a 2000 lecture given by William Winkler, William Yancey and Edward Porter at the U.S. Census Bureau


• Examples are all purely fictitious, but inspired from true cases presented in the above lecture, in Abowd & Vilhuber (2004).