Introduction to Record Linking

John M. Abowd and Lars Vilhuber
April 2013, 2016

Overview

- Introduction to record linking

  What is record linking, what is it not, what is the theory?

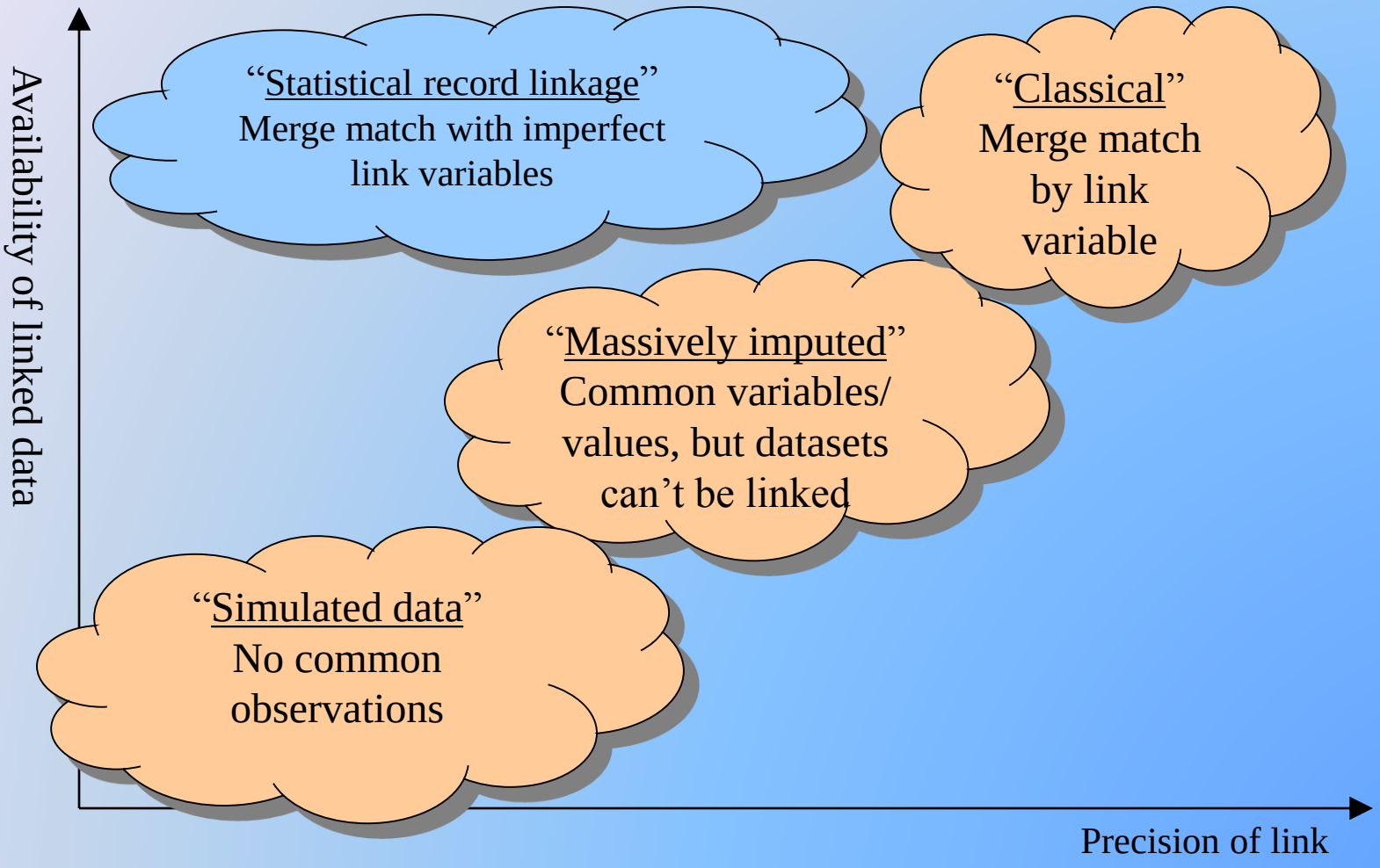
- Record linking: applications and examples

  How do you do it, what do you need, what are the possible complications?

- Examples of record linking

  Do it yourself record linking

From Imputing to Linking



Definitions of Record Linkage

- “A procedure to find pairs of records in two files that represent the same entity”
- “Identify duplicate records within a file”

Uses of Record Linkage

- Merging two files for micro-data analysis
  - CPS base survey to a supplement
  - SIPP interviews to each other
  - Merging years of Business Register
  - Merging two years of CPS
  - Merging financial info to firm survey
- Updating/unduplicating a survey frame or a electoral list
  - Unduplicating the census of population
  - Linking the coverage estimate households to the official enumeration
  - Based on business lists
  - Based on tax records
- Disclosure review of potential public use micro-data

Uses of Record Linkage
(Private-sector Applications)

- Merging two files …
  - Credit scoring
  - Customer lists after merger
  - Internal files when consolidating/upgrading software
- Updating/unduplicating junk mail lists
  - Based on multiple sources of lists
- Disclosure review of potential public use micro-data
  - Not done… (Netflix case)

Types of Record Linkage

- Merging two files for micro-data analysis
  - CPS base survey to a supplement
  - SIPP interviews to each other
  - Merging years of Business Register
  - Merging two years of CPS*
  - Merging financial info to firm survey
- Updating a survey frame or a electoral list
  - Unduplicating the census of population
  - Based on business lists
  - Based on tax records
- Disclosure review of potential public use micro-data

Deterministic linkage: survey-provided IDs

Probabilistic linkage: imperfect or no IDs

Probabilistic linkage: imperfect or no IDs

Probabilistic linkage: no IDs

Need for *Automated* Record Linkage

- RA time required for the following matching tasks:
  - Finding financial records for
    - Fortune 100: *200 hours* (Abowd, 1989)
    - 50,000 small businesses: *??? hours*
  - Identifying miscoded SSNs
    - on 60,000 wage records: several weeks
    - on 500 million wage records: *????*
  - Unduplication of the U.S. Census of Population survey frame (140,000,000 households): *????*
  - Longitudinally linking the 12 million establishments in the Business Register: *????*

Methods of Record Linkage

- Probabilistic record linkage (PBRL)
  - Non-parametric methods
  - Regression-based methods
- Distance-based record linkage (DBRL)
  - Euclidean distance
  - Mahalanobis distance
  - Kernel-based distance

Example: OPM Agencies

Department name	Agency name	Agency type	Employment
DEPARTMENT OF DEFENSE DD	WASHINGTON HEADQUARTERS SERVICES DD21	Cabinet Level Agencies	1481
DEPARTMENT OF DEFENSE DD	OFFICE OF ECONOMIC ADJUSTMENT DD23	Cabinet Level Agencies	38
DEPARTMENT OF AGRICULTURE AG	ECONOMIC RESEARCH SERVICE AG18	Cabinet Level Agencies	388
DEPARTMENT OF AGRICULTURE AG	NATIONAL AGRICULTURAL STATISTICS SERVICE AG20	Cabinet Level Agencies	1084
DEPARTMENT OF AGRICULTURE AG	NATIONAL INSTITUTE OF FOOD AND AGRICULTURE AG22	Cabinet Level Agencies	395
DEPARTMENT OF AGRICULTURE AG	OFFICE OF THE INSPECTOR GENERAL AG23	Cabinet Level Agencies	586
DEPARTMENT OF COMMERCE CM	BUREAU OF THE CENSUS CM63	Cabinet Level Agencies	20987
DEPARTMENT OF COMMERCE CM	OFFICE OF THE INSPECTOR GENERAL CM64	Cabinet Level Agencies	153
DEPARTMENT OF THE TREASURY TR	INTERNAL REVENUE SERVICE TR93	Cabinet Level Agencies	93,654
DEPARTMENT OF VETERANS AFFAIRS VA	OFFICE OF THE SECRETARY VAAA	Cabinet Level Agencies	81
DEPARTMENT OF VETERANS AFFAIRS VA	GENERAL COUNSEL VAAE	Cabinet Level Agencies	758
DEPARTMENT OF VETERANS AFFAIRS VA	INSPECTOR GENERAL VAAF	Cabinet Level Agencies	542

Example: GSA Addresses

National Archives
7th & Pennsylvania Avenue, NW
Washington, DC 20408
Triangle Service Center
Veterans Administration
810 Vermont Avenue, NW
Washington, DC 20420
Potomac Service Center
Department of Education Headquarters
Fourth and C Streets, SW
Washington, DC 20202
DC Service Center
Commerce
14th & Constitution Avenue, NW
Washington, DC 20036
Triangle Service Center

Interior
19th & C Streets, NW
Washington, DC 20240
Potomac Service Center
GSA Headquarters
1800 F Street, NW
Washington, DC 20405
Potomac Service Center
IRS
1111 Constitution Avenue, NW
Washington, DC 20224
Triangle Service Center

Example: Comparing Entities

GSA name	*OPM Department Name*
National Archives	NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
Veterans Administration	DEPARTMENT OF VETERANS AFFAIRS
Department of Education Headquarters	DEPARTMENT OF EDUCATION
Commerce	DEPARTMENT OF COMMERCE
White House	
Interior	DEPARTMENT OF THE INTERIOR
GSA Headquarters	GENERAL SERVICES ADMINISTRATION
IRS	DEPARTMENT OF THE TREASURY

DEFINITIONS

Basic Definitions and Notation

- Entities $a \in A, b \in B$
- Associated files $\alpha(A), \beta(B)$
- Records on files $\alpha(a), \beta(b)$
- Matches

$$M = \{(\alpha(a), \beta(b)) \mid a = b\}$$

- Nonmatches

$$U = \{(\alpha(a), \beta(b)) \mid a \neq b\}$$

Comparisons

- Comparison function maps comparison space into some domain:

$$C : \alpha(a) \times \beta(b) \rightarrow \Gamma$$

- Comparison vector

$$\gamma \in \Gamma, \dim(\gamma) \geq 1$$

- PBRL: Agreement pattern, finitely many values, typically {0,1}, but can be Reals
- DBRL: distance (scalar)

Linkage Rule

- A linkage rule defines a record pair's status based on it's comparison value
  - Link (L)
  - Undecided (Clerical, C)
  - Non-link (N)

$$\mathrm{F}:\Gamma \rightarrow \{L, C, N\}$$

In a Perfect World…

$$(\alpha(a), \beta(b)) \in M \rightarrow (\alpha(a), \beta(b)) \in L$$

and

$$(\alpha(a), \beta(b)) \in U \rightarrow (\alpha(a), \beta(b)) \in N$$

Linkage Rules Depend on Context

- PBRL:
  - *For matching:* rank by agreement ratios, use cutoff values to classify into {L,C,U}
  - *For disclosure-analysis:* rank by agreement ratios, classify as {L} if true link (M) is among top *j* pairs
- DBRL:
  - Rank pairs by distance, link closest pairs

Implementing Probabilistic Record Linkage

- Standardizing
- Blocking and matching variables
- Calculating the agreement index
- Choosing *M* and *U* probabilities
- Estimating *M* and *U* probabilities using EM
- Clerical editing
- Estimating the false match rate
- Estimating the false nonmatch rate

STANDARDIZING

Standardizing

- Standardization is a necessary preprocessing step for all data to be linked via probabilistic record linking
- A standardizer:
  - Parses text fields into logical components (first name, last name; street number, street name, etc.)
  - Standardizes the representation of each parsed field (spelling, numerical range, capitalization, etc.)
- Commercial standardizers have very high value-added compared to home-grown standardizers but are very expensive

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

	Pre	First	Mid	Last	Post1	Post2	Alt1	Std1
1	Dr	William	J	Smith	MD			BWILL
2		Bill		Smith			William	BWILL
3		W	John	Smith	MD			
4		W	J	Smith		Jr		
4		Walter	Jacob	Smith		Sr		WALT

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

	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
1	W	16	Main			St		16		
2				2	215					
3									405	Fuller
4		14588	Hwy	16			W			

Standardizing and Language

- Standardizers are language- and “country”-specific
  - Address tokens may differ: “street”, “rue”, “calle”, “Straße”
  - Address components may differ:
    - 123 Main Street
Normal, IL 61790
    - L 7,1
D-68161 Mannheim
    - 1234 URB LOS OLMOS
PONCE PR 00731-1235

Standardizing and Language (2)

- Names differ
  - Juan, John, Johann, Yohan
- Variations of names differ:
  - Sepp, Zé, Joe -> Joseph
- Frequencies of names differ (will be important later)
  - Juan is frequent in Mexico, infrequent in Germany

Custom Standardization

- Standardization may depend on the particular application
- Example OPM project
  - "Department of Defense"
  - "Department of Commerce"
  - The token "Department of" does not have distinguishing power, but comprises the majority of the "business name"
  - Similar: "Service", "Bureau", "Office of the Inspector General"

PROBABILISTIC RECORD LINKAGE

Example Agreement Pattern

- 3 binary comparisons test whether
  - γ_1 pair agrees on last name
  - γ_2 pair agrees on first name
  - γ_3 pair agrees on street name
- Simple agreement pattern:

  γ=(1,0,1)
- Complex agreement pattern:

  γ=(0.66,0,0.8)

Conditional Probabilities

- Probability that a record pair has agreement pattern γ given that it is a match [nonmatch]

  $P(\gamma|M)$

  $P(\gamma|U)$

- Agreement ratio

  $R(\gamma) = P(\gamma|M) / P(\gamma|U)$

  This ratio will determine the distinguishing power of the comparison γ.

Error Rates

- False match: a linked pair that is not a match (type II error)
- False match rate: probability that a designated link (L) is a nonmatch: $\mu=P(L|U)$
- False nonmatch: a nonlinked pair that is a match (type I error)
- False nonmatch rate: probability that a designated nonlink is a match: $\lambda=P(N|M)$

Fundamental Theorem

1. Order the comparison vectors $\{\gamma^j\}$ by $R(\gamma)$
2. Choose upper T_u and lower T_l cutoff values for $R(\gamma)$
3. Linkage rule:

$$\mathrm{F}:\begin{cases}\gamma_j \in \mathrm{L} & \text{if} \quad R(\gamma_j) \geq T_u \\ \gamma_j \in \mathrm{N} & \text{if} \quad R(\gamma_j) \leq T_l \\ \gamma_j \in \mathrm{C} & \text{else}\end{cases}$$

Fundamental Theorem (cont.)

- Error rates are

$$\mu = \sum_{\gamma^j \in \Gamma} P(\gamma^j \mid U) P(L \mid \gamma^j) = \sum_{\gamma^j \in L} P(\gamma^j \mid U)$$

$$\lambda = \sum_{\gamma^j \in \Gamma} P(\gamma^j \mid M) P(N \mid \gamma^j) = \sum_{\gamma^j \in N} P(\gamma^j \mid M)$$

Fundamental Theorem (3)

- Fellegi and Sunter (*JASA*, 1969):

  If the error rates for the elements of the comparison vector are conditionally independent, then given the overall error rates (μ,λ), the linkage rule *F* minimizes the probability associated with an agreement pattern γ being placed in the clerical review set. (*optimal linkage rule*)

Applying the Theory

- The theory holds on any subset of match pairs (blocks)
- Ratio R: matching weight or total agreement weight
- Optimality of decision rule heavily dependent on the probabilities $P(\gamma \mid M)$ and $P(\gamma \mid U)$

IMPLEMENTING PROBABILISTIC MATCHING

Implementing the Basic Matching Methodology

- Identifying comparison strategies:
  - Which variables to compare
  - String comparator metrics
  - Number comparison algorithms
  - Search and blocking strategies
- Ensuring computational feasibility of the task
  - Choice of software/hardware combination
  - Choice of blocking variables (runtimes quadratic in size of block)
- Estimating necessary parameters

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)

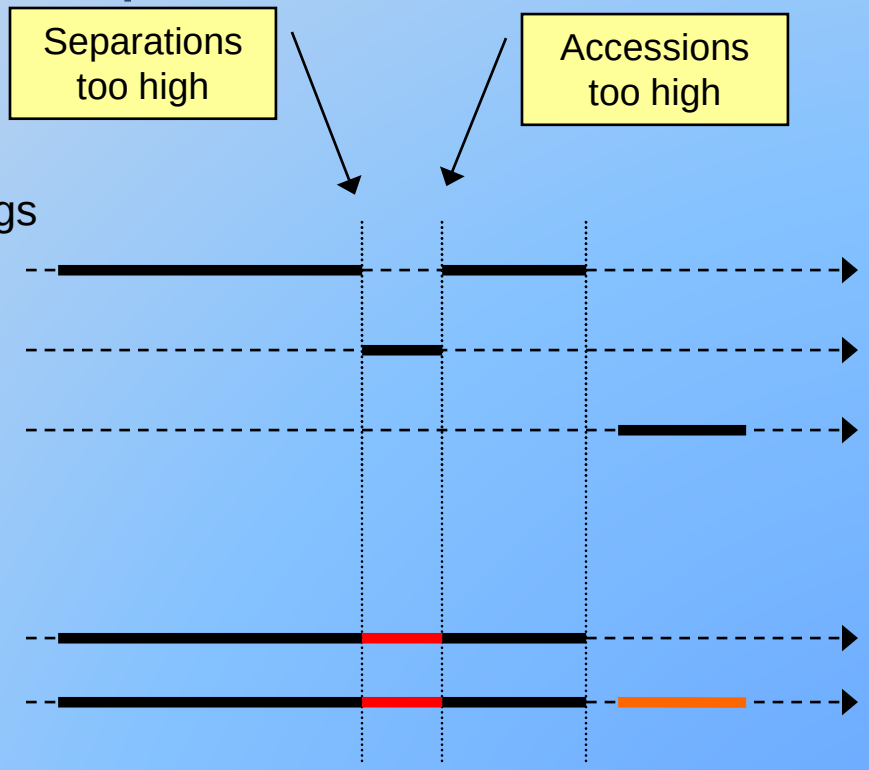
Blocking and Matching

- The essence of a probabilistic record link is iterating passes of the data files in which blocking variables (must match exactly) and matching variables (used to compute the agreement index) change roles.
- Blocking variables reduce the computational burden but increase the false non-match rate => solved by multiple passes
- As records are linked, the linked records are removed from the input files and the analyst can use fewer blocking variables to reduce the false non-matches.
- Matching variables increase the computational burden and manage the tradeoff between false match and false non-match errors

SSN Name Editing

Example

Coded Name	Coded SSN	EIN	Earnings
Lesli Kay	1	A	$10
Leslie Kay	2 1	A	$10
Lesly Kai	3 1	B	$11



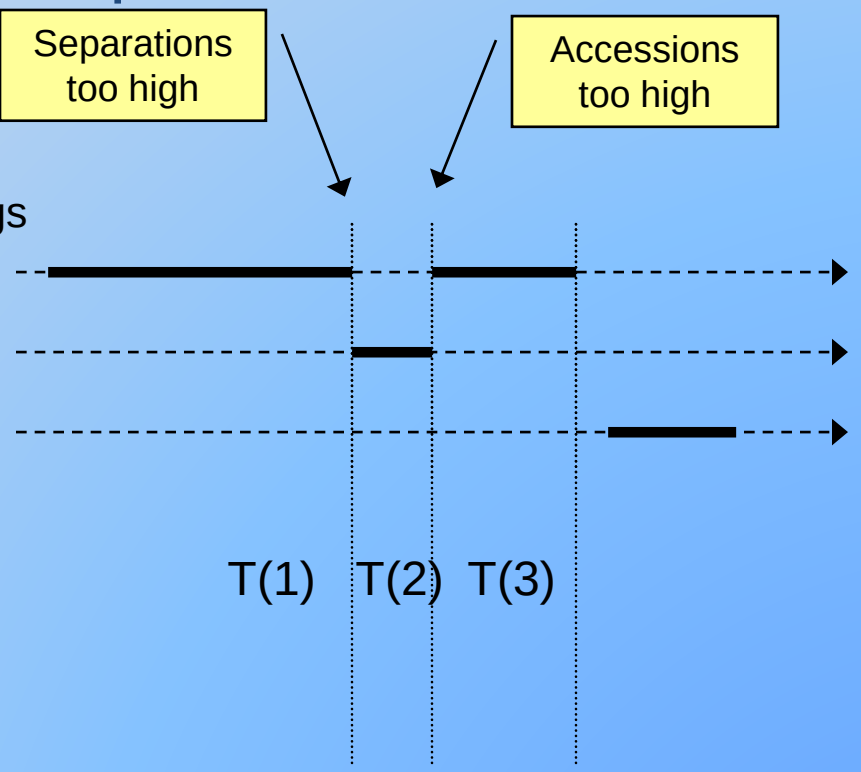
1's tenure with A:

1's employment history

SSN Name Editing

Example

Coded Name	Coded SSN	EIN	Earnings
Lesli Kay	1	A	$10
Leslie Kay	2	A	$10
Lesly Kai	3	B	$11



Computed and observed variables

- Reclassification of information
- Blocking on a-priori information

File	Name	SSN	Earn	Period	Gender	Type
A	Lesli Kay	1	$10	T(2)	M	Hole
B	Leslie Kay	2	$10	T(2)	M	Plug
B	Lesly Kai	3	$10	T(4)	F	Plug

- Blocking: Earn, Period, Gender
- Match on: Name, SSN

Iterating

- First pass may block on a possibly miscoded variable

File	Name	SSN	Earn	Period	Gender	Type
A	Lesli Kay	1	$10	T(2)	M	Hole
B	Leslie Kay	2	$10	T(2)	~~M~~ F	Plug
B	Lesly Kai	3	$10	T(4)	F	Plug

- Block pass 1: Earn, Period, Gender
- Block pass 2: Earn, Period

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)

Soundex: History

- Used for historical analysis, archiving
- Origins in early 20th century
- Available in many computer programs (SQL, SAS, etc.)
- Official "American" Soundex at National Archives: http://www.archives.gov/research/census/soundex.html

String Comparators: Soundex

- The first letter is copied unchanged
- Subsequent letters:

  bfpv -> "1"
  cgjkqsxzç -> "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"

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 variables
- 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(agree|M), P(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(agree|U)* and large samples (population): computed from random matching

Determination of Match Parameters (2)

- Fellegi & Sunter provide a solution when γ 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

Calculating the Agreement Index

- We need to compute $P(\gamma|M)$, $P(\gamma|U)$ and the agreement ratio $R(\gamma) = P(\gamma|M) / P(\gamma|U)$
- The agreement index is ln $R(\gamma)$
- The critical assumption is conditional independence:

  $P(\gamma|M) = P(\gamma_1|M)\, P(\gamma_2|M) \ldots P(\gamma_K|M)$

  $P(\gamma|U) = P(\gamma_1|U)\, P(\gamma_2|U) \ldots P(\gamma_K|U)$

  where the subscript indicates an element of the vector γ.
- Implies that the agreement index can be written as:

$$\ln R(\gamma) = \sum_{k=1}^{K} \ln\left(\frac{P(\gamma_k|M)}{P(\gamma_k|U)}\right)$$

Choosing *m* and *u* Probabilities

- Define

  $m_k = P(\gamma_k | M)$

  $u_k = P(\gamma_k | U)$

- These probabilities are often assessed using *a priori* information or estimated from an expensive clerically edited link.
  - *m* often set *a priori* to 0.9
  - *u* often set *a priori* to 0.1
- Neither of these assumptions has much empirical support

Some Rules of Thumb

- Gender

  $m_k = P(\gamma_k|M)$ is a function of the data (random miscodes of gender variable)

  $u_k = P(\gamma_k|U) = 0.5$ (unconditional on other variables). This may not be true for certain blocking variables: age, veteran status, etc. will affect this value

- Exact identifiers (SSN, SIN)

  $m_k = P(\gamma_k|M)$ will depend on verification by the data provider. For example, embedded checksums will move this probability closer to 1

  $u_k = P(\gamma_k|U) << 0.1$

Marginal Probabilities: Educated Guesses for *Starting* Values

- P(agree on characteristic X| M)=
  - 0.9 if X = first, last name, age
  - 0.8 if X = house no., street name, other characteristic
- P(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(first)=0.9/0.1=9) is larger than the street name (R(street)=0.8/0.2=4)

Marginal Probabilities: Better Estimates of P(agree|M)

- P(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(agree|M) based on known truth sample

Estimating *m* and *u* Using Matched Data

- If you have two files α and β that have already been linked (perhaps clerically, perhaps with an exact link) then these estimates are available:

$$\hat{m}_k = \frac{\sum_{(a,b)\in L} \gamma_k(a,b) = 1}{\sum_{\forall(a,b)} 1\left[(a,b) \in L\right]}$$

$$\hat{u}_k = \frac{\sum_{(a,b)\in U} \gamma_k(a,b) = 1}{\sum_{\forall(a,b)} 1\left[(a,b) \in U\right]}$$

$$\text{where } a \in \alpha, b \in \beta, \gamma(a,b) \in \Gamma$$

Estimating *m* and *u* Probabilities Using EM

- Based on Winkler 1988 "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671
- Uses the identity

$$P(\gamma)=P(\gamma|M)P(M)+P(\gamma|U)P(U)$$

- Imposes conditional independence

Clerical Editing

- Once the *m* and *u* probabilities have been estimated, cutoffs for the U, C, and L sets must be determined
- This is usually done by setting preliminary cutoffs then clerically refining them
- Often the *m* and *u* probabilities are tweaked as a part of this clerical review

Estimating the False Match Rate

- This is usually done by clerical review of a run of the automated matcher
- Some help is available from Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694-707

Estimating the False Nonmatch Rate

- This is much harder
- Often done by a clerical review of a sample of the non-match records
- Since false nonmatching is relatively rare among the nonmatch pairs, this sample is often stratified by variables known to affect the match rate
- Stratifying by the agreement index is a very effective way to estimate false nonmatch rates

Post-processing

- Once matching software has identified matches, further processing may be needed:
  - Clean up
  - Carrying forward matching information
  - Reports on match rates

DISTANCE-BASED RECORD LINKING

Distance-Based Record Linking

- Distance between any pair of records

$$(\alpha(a), \beta(b)) = (\alpha, \beta)$$

can be generally defined as

$$d(\alpha, \beta)^2 = (\alpha - \beta)'\big(Var(A) + Var(B) - 2Cov(A, B)\big)^{-1}(\alpha - \beta)$$

DBRL: 4 Cases

- Mahalanobis distance, known covariance

$$Cov(A, B)$$

- Mahalanobis distance, unknown covariance

$$Cov(A, B) = 0$$

DBRL: 4 Cases

- Euclidean distance, unstandardized inputs

$$Var(A)+Var(B)-2Cov(A,B)=I$$

- Euclidean distance, standardized inputs

$$\tilde{F}=\frac{F-\bar{F}}{\sqrt{Var(F)}}\sim N(0,1)$$

$$Var(\tilde{A})+Var(\tilde{B})-2Cov(\tilde{A},\tilde{B})=I$$

Linkage Rules

- *Matching:* Sort by distance, choose top *j* pairs as matches
- *Disclosure analysis:* Sort by distance, identify true matches among top *j* pairs

Multi-file matching

- As the number of sources increases (n-file matching, n>>2), **clustering/ machine-learning** kicks in
- Grouping together similar “records” based on a distance metric
- Works better when many instances of an entity exist
- Hard (NP-hard)

Examples

- Sadinle and Fienberg (JASA 2013)
- "Methods for Quantifying Conflict Casualties in Syria", Steorts, Ventura, Sadinle, Fienberg (all CMU at the time), JSM 2015
- Watch the 3-min explanation at http://events.technologyreview.com/emtech/15/video/watch/innovators-under-35-rebecca-steorts/

EXAMPLES

Example: Abowd and Vilhuber (2005)

- "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business and Economic Statistics*, 23(2), pages 133-152
- *Goal of the study*: assess the impact of measurement error in tenure on aggregate measures of turnover
- http://www.jstor.org/stable/27638803
- Appendix A has a detailed description of matching passes and construction of additional blocking variables

Example: CPS-Decennial 2000 match

- "Accuracy of Data for Employment Status as Measured by the CPS-Census 2000 Match," Palumbo and Siegel (2004), *Census 2000 Evaluation B.7*, accessed at http://www.census.gov/pred/www/rpts/B.7%20Final%20Report.pdf on April 7, 2013
- *Goal of the study*: assess employment status on two independent questionnaires

Example: SIPP Synthetic Beta

- "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project," John M. Abowd, Martha Stinson and Gary Benedetto (2006), http://www.census.gov/sipp/FinalReporttoSocialSecurityAdministration.pdf, on April 7, 2013
- *Goal of the study*: assessing disclosure risk

SSB: Probabilistic Record Linkage

- Blocking strategy
  - Unsynthesized variables: *male* and *maritalstat*
  - Segments of *personid* (confidential variable)
    - An "intruder" would not have access to personid, would need to compare more records within blocks
    - Matching results for the purpose of disclosure analysis is thus a conservative estimates: match rates are likely to be lower when **not** using *personid*.
- Matching variables
  - Hispanic, education, disability, number of children, industry and occupation of job, place of birth, whether owned a home, pension information

SSB: Assessing Match Rates

- Has access to original data – match rates can be computed
  - *m* and *u* computed exactly and used in the matching procedure
  - Again, more information is used than available to an intruder
- Classify by total agreement weight, allow for any match, keep the best two matches (greedy algorithm)

SSB: Match Rates and Re-Identification

- For married people, less than 1%
- For single people, no more than 1.2%
- Using Distance-based matching,
  - the highest re-identification rate is 2.91% when using a known variance-covariance matrix and for certain subgroups (this is not tabulated separately for PBRL)
  - Average across all groups when covariance matrix is unknown is 0.75% and when covariance matrix is I, 1.2%

FINAL NOTES ON SOFTWARE

Matching Software

- Commercial ($$$$-$$$$$)
  - Automatch/Vality/Ascential/IBM WebSphere Information Integration
    (grew out of Jaro's work at the Census Bureau)
  - DataFlux/ SAS Data Quality Server
  - Oracle
  - Others
- Custom software (0-$$)
  - C/Fortran Census SRD-maintained software
  - Java implementation used in Domingo-Ferrer, Abowd, and Torra (2006)
  - Java Data Mining API

Specific other tools

- RecordLinkage R package, see http://cran.r-project.org/web/views/NaturalLanguageProcessing.html
- "STATA utilities to facilitate probabilitistic record linkage methods" Nada Wasi and Aaron Flaaen (Michigan), http://www.ncrn.info/software/stata-utilities-facilitate-probabilitistic-record-linkage-methods

Software Differences

- Each software is an empirical/practical implementation driven by specific needs
- Terminology tends to differ:
  - “standardize”, “schema”, “simplify”
  - “block”, “exact match”
  - “comparator function”, “match definition”

	Custom/ SRD	*Vality/Ascential*	*SAS DQ*
Standardizing	`standard`	Yes	`PROC DQSCHEME`
Blocking and matching variables	Yes	Yes	Yes (blocking = "exact match")
Calculating the agreement index	Yes	Yes	No
Choosing *M* and *U* probabilities	Yes	Yes	No (varying "sensitivity" achieves the same goal)
Estimating *M* and *U* probabilities using EM	Yes (`eci`)	(external)	No
Matching	`matcher`	Yes	`PROC DQMATCH`

Software Notes

- Idiosyncratic implementations
  - Merge two files in Stata/SAS/SQL/etc. (outer join or wide file), within blocks
  - For given m/u, compute agreement index on a per-variable basis for all
  - Order records within blocks, select match record
  - Generically, this works with numeric variables, but has issues when working with character variables
    - Typically, only Soundex available
    - Implementing string comparators (Jaro) may be difficult

Software notes: character matching

- SAS Data Quality
  - Is focused on “data quality”, “data cleaning”
  - Has functions for standardization and string matching
  - String matching is variant of Soundex: simplified strings based on generic algorithms
  - *Issues*: no probabilistic foundation, useful primarily for sophisticated string matching
- Combination of SAS DQ and idiosyncratic processing can yield powerful results

Software Notes: SQL

- SQL languages are used in many contexts where matching is frequent
  - Oracle 10g R1 has Jaro-Winkler edit-distance implemented.
  - MySQL allows for some integration (possibly see http://androidaddicted.wordpress.com/2010/06/01/jaro-winkler-sql-code/)
  - PostGreSQL add-ons http://pgsimilarity.projects.postgresql.org/

Software notes: R

- R has a recent package (untested by us) that seems a fairly complete suite

  http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf

- Does not include standardization

Software notes: Stata

- Stata has no native matching, but user-provided tools
  - reclink
  - "STATA utilities to facilitate probabilitistic record linkage methods" Nada Wasi and Aaron Flaaen (Michigan), http://www.ncrn.info/software/stata-utilities-facilitate-probabilitistic-record-linkage-methods , specifically focused on standardizing of firm names, but also general purpose record linking tools
- SAS version of the custom standardizer exists as well

Acknowledgements

- This lecture is based in part on a 2000 and 2004 lecture given by William Winkler, William Yancey and Edward Porter at the U.S. Census Bureau
- Some portions draw on Winkler (1995), "Matching and Record Linkage," in B.G. Cox et. al. (ed.), *Business Survey Methods*, New York, J. Wiley, 355-384
- Some (non-confidential) portions drawn from Abowd, Stinson, Benedetto (2006), "Final Report to Social Security Administration on the SIPP/SSA/IRS Public Use File Project"