An Introduction to Probabilistic Record Linkage

John ‘Mac’ McDonald

Centre for Longitudinal Studies

Institute of Education, London

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

ESRC National Centre for
Research
Methods

IOE
LONDON

Leading education
and social research
Institute of Education
University of London

A·D·M·I·N
Administrative Data -
Methods, Inference & Network

CLS

Record linkage

Record linkage (also known as data linkage)

- □ **for organising ONE dataset**
  - – data cleaning
  - – removing duplicates
- □ **for merging TWO OR MORE datasets**
  - – merging individual-level datasets
  - – adding census data to survey data

Identification of Duplicates Given Name, Address, Age

Matching Information

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17
Javier Martinez	49 E Applecross Road	33
Haveir Marteenez	49 Aplecross Raod	36
Gillian Jones	645 Reading Aev	22
Jilliam Brown	123 Norcross Blvd	43

Record linkage . . .

“[is] a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events (said to be matched).”

Fellegi IP & Sunter AB (1969) A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210

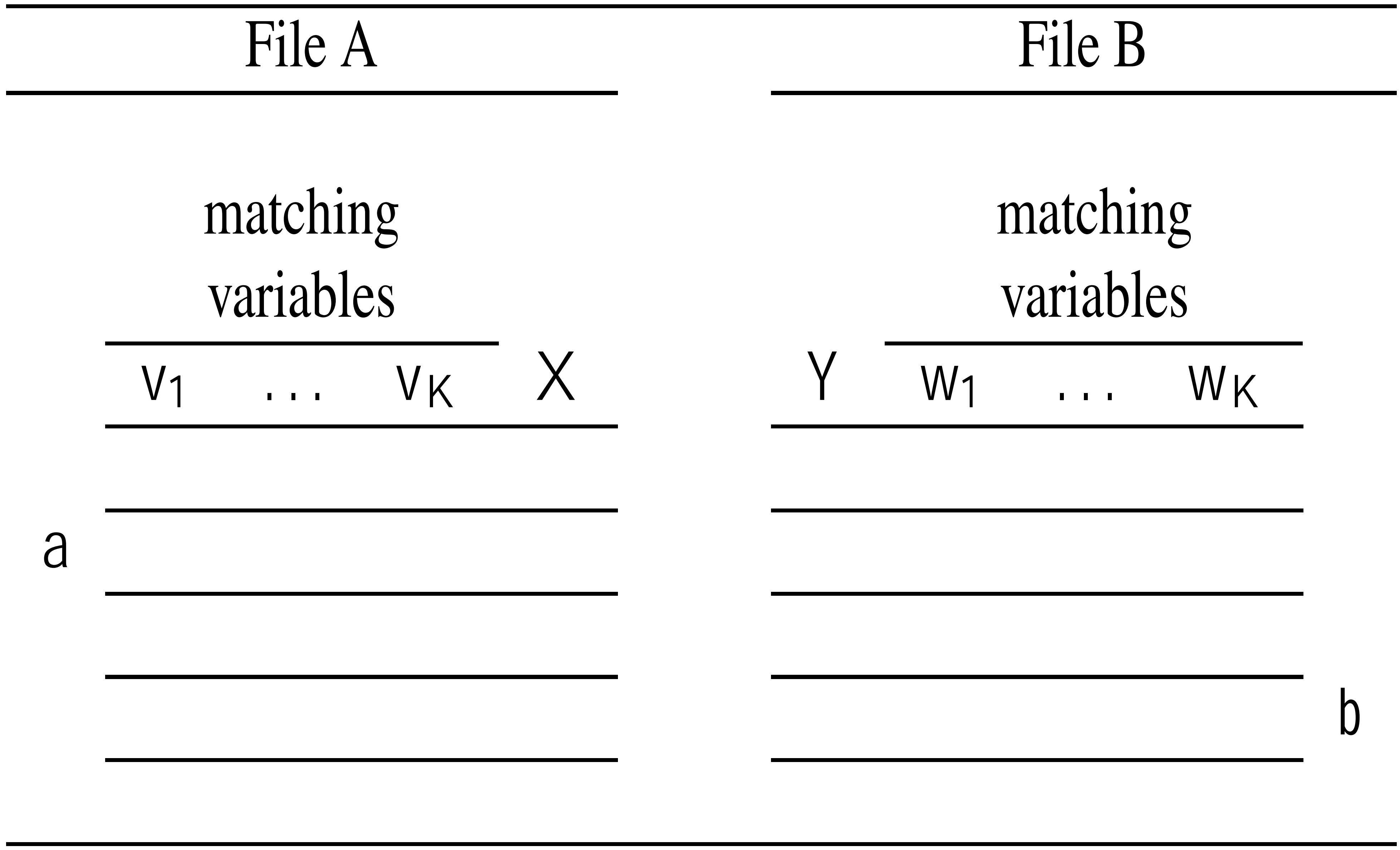
Problem of record linkage

problem - **quickly and accurately** determining if pairs of records describe the same entity, but unique IDs to bring together the matching records are lacking

records must contain some common identifying information (**keys or matching variables**)

- □ unique identifier (ideal in theory)
- □ name and/or address
- □ age (DOB) and sex

Files A & B, record a in A & record b in B



File A

File B

matching variables

matching variables

$v_1 \quad \ldots \quad v_K$ $\quad X$

$Y \quad$ $w_1 \quad \ldots \quad w_K$

a

b

Methodology of record linkage

□ **two distinct methodologies for data linkage**

□ **deterministic linkage** methods involve exact one-to-one character matching of linkage variable(s)

□ **probabilistic linkage** methods involve the calculation of linkage weights estimated given all the observed agreements and disagreements of the data values of the matching variable(s)

□ **probabilistic linkage methods can lead to much better linkage than simple deterministic linkage methods**

Deterministic linkage

- □ **simplest method of matching** - sort/merge
- □ exact matching ONLY works well if the linking data are perfect and present in all the databases you want to link
- □ works best when there is a single unique identifier (key)
- □ otherwise, matching based on sets of identifiers predetermined by the researcher
- □ identifiers have equal weight
- □ identifiers chosen by researcher or by availability
- □ **works best with high quality data, but yields less success than probabilistic linkage**

Deterministic linkage . . .

□ **deterministic matching links records**

- using a fixed set of matching variables
- **exact 1-to-1 character matching**

□ **problems**

- **often no unique, known and accurate ID**
- **missing values & partial agreements common**

□ sometimes only the first few characters of a field are used with a wildcard substituted for later characters

- Anders*, for Anderson and Andersen
- Martin*, but Martin and Martinez also match

Data linkage . . .

Data linkage is a challenging problem because of

- □ errors, variations and missing data on the information used to link records
- □ differences in data captured and maintained by different databases, e.g. age versus DOB
- □ data dynamics and database (DB) dynamics as data regularly and routinely change over time
  - – name changes due to marriage & divorce
  - – address changes

Data problems

□ typos/mispelling

□ letters or words out of order

□ fused or split words

□ missing or extra letters

□ incomplete words

□ extraneous information

□ incorrect or missing punctuation

□ abbreviations

□ multiple errors

Surname	Name	Day of B	Year of B	freq
0	0	0	0	414138
0	0	0	1	5321
0	0	1	0	14004
0	0	1	1	168
0	1	0	0	3090
0	1	0	1	43
0	1	1	0	102
0	1	1	1	9
1	0	0	0	969
1	0	0	1	17
1	0	1	0	22
1	0	1	1	19
1	1	0	0	14
1	1	0	1	9
1	1	1	0	6
1	1	1	1	513

Linkage projects typically have three phases

□ **pre-linkage**

- data cleaning
- processing data fields to recognize similarity

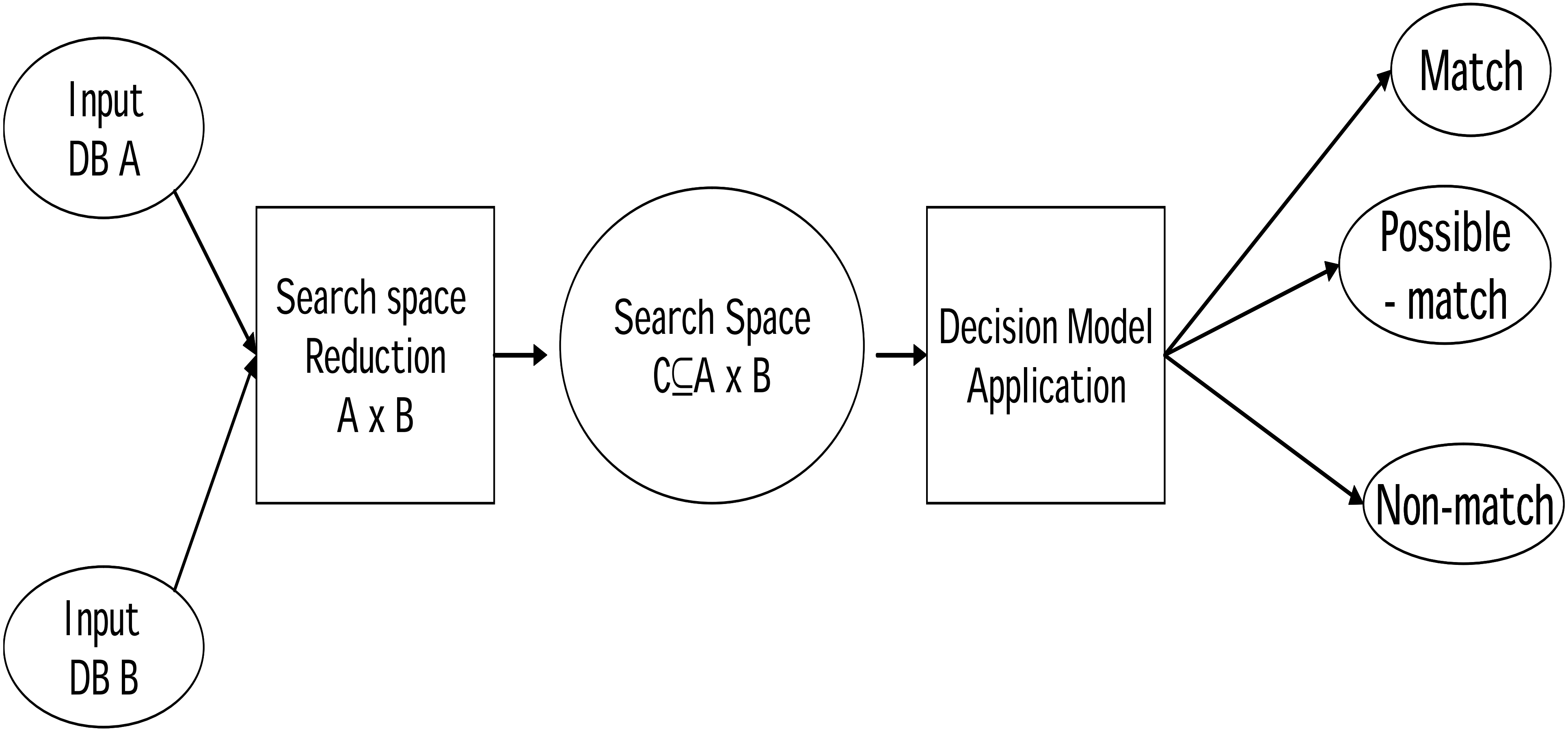
□ **linkage phase: deciding whether two records are a**

- **duplicate**
- **match (link)**

□ **post-linkage**

- manual/clerical review of unlinked records
- research using the linked data

Phases of record linkage



Hypothetical example

- □ Clark DE (2004) Practical introduction to record linkage for injury research, Injury Prevention, 10, 186-191
- □ $N_A = 10$ **ambulance cases**
- □ $N_B = 20$ **emergency department cases**
- □ **aim to match the cases**
- □ **prior belief - anticipated number of correct matching records is** $N_X = 9$

http://injuryprevention.bmj.com/content/10/3/186.full.pdf

Prior probability of a match

□ **prior probability** that a randomly selected record from file A matches a randomly selected record from file B

$$\Pr(\text{match}) = \frac{N_X}{N_A} \times \frac{1}{N_B} = \frac{9}{10} \times \frac{1}{20} = \mathbf{0.045}$$

□ generally, this probability will be a very small number, so the prior odds will be similar

□ **prior odds: 0.045 / (1 - 0.045) = 0.047**

□ usually we will work with log odds of a match, which will be a negative number

File A - Ambulance data

Case	Year	Day	Hosp	Birth Year	Birthday	Sex
A01	01	Jan01	X	1950	Jan21	M
A02	01	Jan01	X	1950	May01	F
A03	01	Jan10	Y	1975	Dec27	
A04	01	Aug13	X	1977	Apr29	F
A05	01	Sep12	Y	1980	Feb16	F
A06	01	Dec31	Z	1919	Sep16	M
A07	02	Feb02	X	1924	Mar26	F
A08	02	Jun10	Y	1951	Mar29	M
A09	02	Aug06	Y	1953	Apr17	
A10	02	Sep21	Z	1956	Jun03	F

File B - Emergency department data

Case	Year	Day	Hosp	Birth Year	Birthday	Sex
E01	01	Jan01	X	1950	Jan21	M
E02	01	Jan10	Z	1987	Jul17	M
E03	01	Feb23	X	1992	Oct19	M
E04	01	Apr22	Y	1979	May09	M
E05	01	May02	X	1929	Nov12	F
E06	01	May23	Y	1964	Dec01	M
E07	01	Jun01	X	1950	May01	F
E08	01	Aug14	X	1977	Apr29	F
E09	01	Sep12	Y	1980	Feb16	F
E10	01	Oct21	Y	1985	Mar12	M

File B - Emergency department data . . .

Case	Year	Day	Hosp	Birth Year	Birthday	Sex
E11	02	Jan01	Z	1919	Sep16	M
E12	02	Jan10	Y	1975	Dec27	F
E13	02	Feb02	X	1924	Mar26	F
E14	02	May16	X	1924	Oct12	M
E15	02	Jun10	Y	1951	Mar29	M
E16	02	Jul04	Z	1982	Jun12	M
E17	02	Aug05	Y	1953	Apr17	M
E18	02	Aug06	Y	2002	Apr17	F
E19	02	Sep21	Z	1956	Jun03	F
E20	02	Nov22	X	1917	May29	M

Comparing record pairs - exact matches

Case	Year	Day	Hosp	Birth Year	Birthday	Sex
A10	02	Sep21	Z	1956	Jun03	F
E19	02	Sep21	Z	1956	Jun03	F
A01	01	Jan01	X	1950	Jan21	M
E01	01	Jan01	X	1950	Jan21	M
A05	01	Sep12	Y	1980	Feb16	F
E09	01	Sep12	Y	1980	Feb16	F
A07	02	Feb02	X	1924	Mar26	F
E13	02	Feb02	X	1924	Mar26	F
A08	02	Jun10	Y	1951	Mar29	M
E15	02	Jun10	Y	1951	Mar29	M

Comparing record pairs - matches?

Case	Year	Day	Hosp	Birth Year	Birthday	Sex
A03	01	Jan10	Y	1975	Dec27	
E12	02	Jan10	Y	1975	Dec27	F
A02	01	Jan01	X	1950	May01	F
E07	01	Jun01	X	1950	May01	F
A04	01	Aug13	X	1977	Apr29	F
E08	01	Aug14	X	1977	Apr29	F
A09	02	Aug06	Y	1953	Apr17	
E17	02	Aug05	Y	1953	Apr17	M
E18	02	Aug06	Y	2002	Apr17	F
A06	01	Dec31	Z	1919	Sep16	M
E11	02	Jan01	Z	1919	Sep16	M

M-probability (Match probability)

- □ **M-probability: probability that a field agrees given that the pair of records is a true match**

- □ **for any given field, the same M-probability applies for all records**

- □ **assume the following:**
  - – admission year: .99
  - – admission date: .95
  - – hospital: .99
  - – birth year: .95
  - – birthday: .99
  - – sex: .95

M-probability (Match probability) . . .

□ **data quality is quantified by the M-probabilities**

□ M-probability of 0.95 for surname means that the probability two records belonging to the same person will agree on last name is 0.95

□ why will surname on two records belonging to the same person disagree 5% of the time

- data entry errors
- missing data
- instability of value, e.g. surname change
- misspelling, e.g. Anderson versus Andersen

U-probability (Unmatch probability)

- □ **U-probability: probability that a field agrees given that the pair of records is NOT a true match**

- □ **often simplified as the chance that 2 records will randomly match**, i.e. proportion of records with a specific value on the larger file

- □ **the U-probability is defined as value specific and will often have multiple values for each field**

- □ assume:
  - admission year: .5; admission date: .0027 (1/365)
  - **hospital X or Y: .4 hospital Z: .2**
  - birth year: .01 (1/100); birth date: .0027 (1/365)
  - **males: .6; females: .4**

U-probability (Unmatch probability) . . .

□ **probability of random matches is context specific**

□ **generally, gender is of limited value for linkage**

□ randomly agrees 50% of the time in most contexts, but

- males: .6; females: .4 for ambulance data
- males: .0; females: 1.0 in an all female school

□ **gender in an all female school is useless for linkage as one would obtain a match on that field in any random pairing**

U-probability (Unmatch probability) . . .

- □ **consider a field with a unique value for every person in the dataset**, e.g. unique ID number
  - **this field can be very useful for linkage if it is in both files**
  - one would not expect to obtain a match randomly
  - only limiting factor on correct matches would be the data quality, the value of the M-probability

Jenkins S et al (2006) **The feasibility of linking household survey and administrative record data**: New evidence for Britain. International Journal of Social Research Methodology 11, 29-43

U-probability (Unmatch probability) . . .

- □ **consider the field surname**
- □ **different people will have different U-probabilities**, depending on their own specific surname (and context, e.g. country/region)
- □ **how are U-probabilities estimated?**
- □ typically estimated as the proportion of records with a specific value, based on the frequencies in the primary or more comprehensive and accurate data source

U-probability (Unmatch probability) . . .

□ **how are U-probabilities estimated for surnames?**

□ 300 000 birth certificates

□ 600 Andersons

□ 30 Rumplestilskins

□ estimated U-probability

  – 600 / 300 000 = .0020 for Anderson

  – 30 / 300 000 = .0001 for Rumplestilskin

Estimating match (probabilistic) weights

- **for a given field with match probability M and unmatch probability U**

- **for an agreement, we calculate the weight**

$$\log(\frac{M}{U})$$

- **for a disagreement, we calculate the weight**

$$\log(\frac{1-M}{1-U})$$

- assuming independence of information across the fields, **we sum the weights across all the fields** to obtain the total weight for the record pair

Estimating match (probabilistic) weights

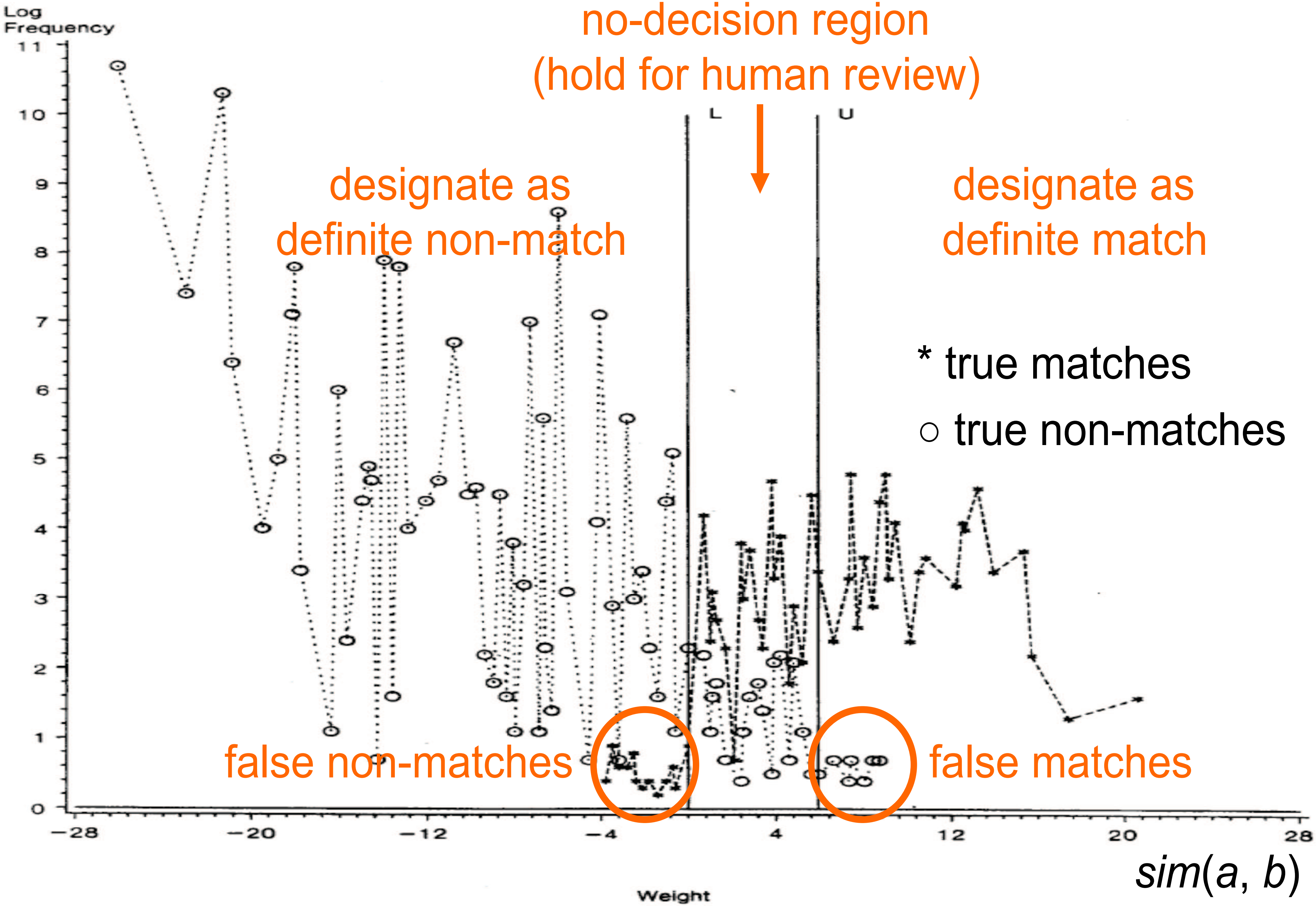
- □ **highest weight 7.455 is for the pair A10-E19, which agrees across all fields**, as calculated by

$$\log(\frac{.99}{.50})+\log(\frac{.95}{.0027})+\log(\frac{.99}{\mathbf{.20}})+\log(\frac{.95}{.01})+\log(\frac{.99}{.0027})+\log(\frac{.95}{\mathbf{.4}})$$

- □ **for pair A03-E12, the admission year and sex were different, and the weight is 4.704**, as calculated by

$$\log(\frac{\mathbf{1-.99}}{\mathbf{1-.50}})+\log(\frac{.95}{.0027})+\log(\frac{.99}{.20})+\log(\frac{.95}{.01})+\log(\frac{.99}{.0027})+\log(\frac{\mathbf{1-.95}}{\mathbf{1-.4}})$$

Fellegi-Sunter model



no-decision region
(hold for human review)

designate as
definite non-match

designate as
definite match

* true matches

○ true non-matches

false non-matches

false matches

sim(*a*, *b*)

Posterior odds and posterior probabilities

- □ **posterior odds = prior odds × likelihood**

- □ **for pair A10-E19, the posterior odds are 1 340 000** as calculated from

$$.047 \times \frac{.99}{.50} \times \frac{.95}{.0027} \times \frac{.99}{.20} \times \frac{.95}{.01} \times \frac{.99}{.0027} \times \frac{.95}{.4}$$

- □ **for pair A03-E12, admission year and sex differed, and the posterior odds are 2 376** as calculated from

$$.047 \times \frac{1-.99}{1-.50} \times \frac{.95}{.0027} \times \frac{.99}{.20} \times \frac{.95}{.01} \times \frac{.99}{.0027} \times \frac{1-.95}{1-.4}$$

Posterior odds and posterior probabilities . . .

□ posterior probability $= \dfrac{\text{posterior odds}}{1 + \text{posterior odds}}$

□ for pair A10-E19, the posterior probability is .9999

□ for pair A03-E12, admission year and sex differed, and the posterior probability is .9996

□ A09 is problematic because it might be matched to either E17 or to E18 with posterior probabilities of .9805 or .9921 respectively

□ pair A06-E11 is also uncertain with posterior probability of .9507 as pair differs by year and day, but Dec 31, 2001 and Jan 01, 2002!

Probabilistic linkage

- **each matching variable is compared and assigned a score (weight) based on how well it matches**
- frequency analysis of data values is important
- uncommon value agreement stronger evidence for linkage, e.g. Rumplestilskin versus Smith
- **calculates a score for each field that indicates, for any pair of records, how likely it is that they both refer to the same entity**
- **sum the scores over fields**
- sort record pairs in order of their scores (weights)

Probabilistic linkage . . .

- cut off values for scores (weights) are used to distinguish between matches and non-matches
- **above a certain threshold, everything is a match** (link)
- **below a certain threshold, nothing is a match** (nonmatch or nonlink)
- **in between (grey area), possible match needs manual/clerical review**

Probabilistic linkage . . .

- □ **total score for a link between any two records is the sum of the scores generated from matching individual fields**

- □ score assigned to a matching of individual fields
  - – is based on the probability that a matching variable agrees given that a comparison pair is a match
  - – M-probability - similar to “sensitivity”, i.e. the proportion of actual positives which are correctly identified

Probabilistic linkage . . .

- □ score assigned to a matching of individual fields
  - – reduced by the probability that a matching variable agrees given that a comparison pair is not a match (U = unmatched)
  - – U-probability - similar to “specificity”, i.e. the proportion of negatives which are correctly identified
- □ **agreement argues for linkage**
- □ **disagreement argues against linkage**
- □ full agreement stronger evidence for linkage than partial agreement

Probabilistic linkage . . .

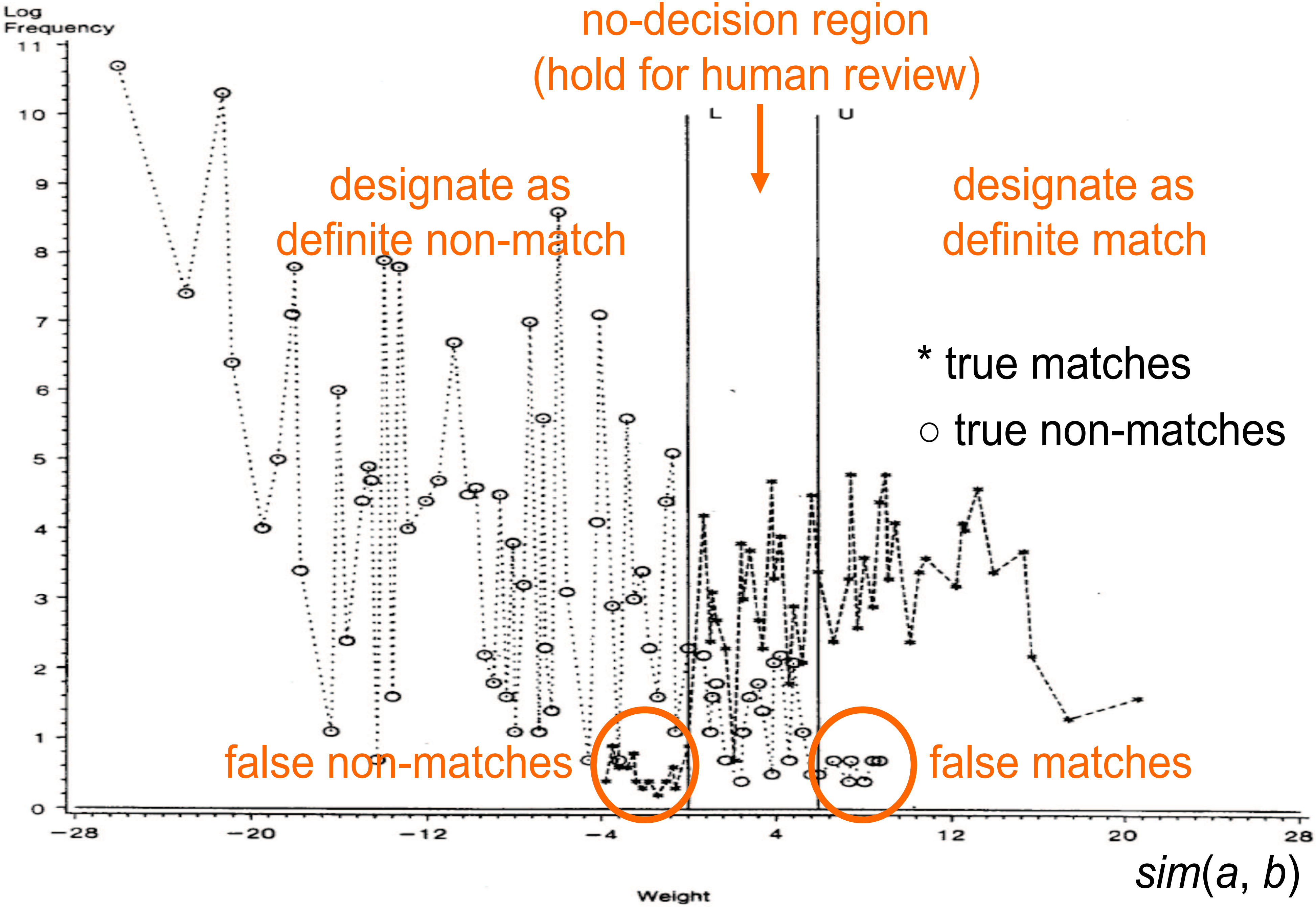
□ based on the probabilities of agreement or disagreement between the identifiers

□ all identifiers do not have equal weight

□ **accurate linkage is mainly dependent on the amount of discriminating power inherent in the variables common to the records that need to be matched and 'good' data**

Fellegi IP & Sunter AB (1969) A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210

Fellegi-Sunter model



no-decision region
(hold for human review)

designate as
definite non-match

designate as
definite match

* true matches

○ true non-matches

false non-matches

false matches

sim(*a*, *b*)

Deterministic versus probabilistic methods

studies using human review or artificially withheld identifying information as 'gold standard'

Gomatam S et al (2002) An empirical comparison of record linkage procedures, Statistics in Medicine, 21, 1485-1496

http://nisla05.niss.org/dgii/presentations/gomatam-sic-200305.pdf

Roos LL, Walld R, Wajda A, Bond R, Hartford K (1996) Record linkage strategies, outpatient procedures, and administrative data. Medical Care 34, 570–582

Jamieson E, Roberts J, Browne G (1995) The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. Methods Inf Med. 34, 371–377

RELAIS

□ RELAIS (Record Linkage At IStat) toolkit

□ **an open source toolkit for building record linkage workflows**

□ JAVA based

□ statistical methods implemented in R

http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/

References

Herzog TN, Scheuren FJ & Winkler WE (2007). Data quality and record linkage techniques. New York: Springer. 234 pp. Part IV discusses software.

Howe GR (1998) Use of computerized record linkage in cohort studies. Epidemiologic Reviews 20, 112–121

Krewski D, Dewanji A, Wang Y, Bartlett S, Zielinski JM & Mallick R (2005) The effect of record linkage errors on risk estimates in cohort mortality studies. Survey Methodology 31,13–21

Brenner H, Schmidtmann I & Stegmaier C (1997) Effect of record linkage errors on registry-based follow-up studies. Statistics in Medicine 16, 2633–2643