

# Overview of Record Linkage for Name Matching

W. E. Winkler, [william.e.winkler@census.gov](mailto:william.e.winkler@census.gov)

NSF Workshop, February 29, 2008

## Outline

1. Components of matching process and nuances

## Match NSF file of Ph.D. recipients against alternate files

**Goal:** Determine how many patents, refereed papers, etc. are associated with each Ph.D.

1. Determine common information in two files that can be used for matching. *First name* (or initial) and *Last name*; other information.
2. Extract common information and put in form for comparison.

Free-form	First	Formatted Middle Init	Last
John H. Smith	John	H	Smith
Smith, J. H.	J	H	Smith

Free-form	Formatted			
		First	Middle Init	Last
J. H. Smith and M. Q. Jones	r1a	J	H	Smith
	r1b	M	Q	Jones
Smith, J. H., and Jones, M.Q.	r2a	J	H	Smith
	r2b	M	Q	Jones

Need software that can locate name information (if not in specific location) and to do reformatting as above

3. Need to determine other fields that can be used in matching  
 ?????

I.e., is there information beyond last name and first initial?

3 million 'Smith's – first initial plus last name is not sufficient

#### 4. **Typographical Variation: Need approximate string comparison**

*Possible true match*

‘J. A. Smith’ versus ‘J. H. Smoth’

Is ‘A’ really ‘H’?    Is ‘Smith’ really ‘Smoth’?

*Similar but likely nonmatch*

‘Mari M. Jones’ versus ‘Marvin N. Janes’

*Edit distance* – count minimum number of insertions, deletions, substitutions to get from one string to another

*Jaro-Winkler string comparator* – 10 times as fast as edit distance; in most elementary applications works as well

## 5. Searching and Retrieval in Large Files

60,000 in File A against File B of 3 million

*Blocking*: Only bring together pairs of records that agree on last name (or other characteristics if available).

Reduce number of comparisons of records from

$$1.8 \times 10^{11} = 6 \times 10^4 \text{ times } 3 \times 10^6 \text{ to}$$

$$4.0 \times 10^8$$

CPU time:  $10^5$  pairs per second yields 3-5 hours

## 6. Improve parameter estimation in record linkage decision rules

Fellegi and Sunter (1969 *JASA*) **Model of Record Linkage**

Began with ideas of Newcombe (*Science* 1959, *CACM* 1962)

Let  $A$  and  $B$  be two files. Let  $\gamma$  be an arbitrary agreement pattern in a comparison space  $\Gamma$  on  $A \times B = M \cup U$  where  $M$  are matches and  $U$  are nonmatches.

$$R = P(\gamma | M) / P(\gamma | U) \quad (1)$$

$\gamma$  could be agree/disagree on fields such as first name, last name, etc.

could be partial agrees

could account for relative frequency – ‘Zabrinsky’ rarer than ‘Smith’

The *classification rule* is given by:

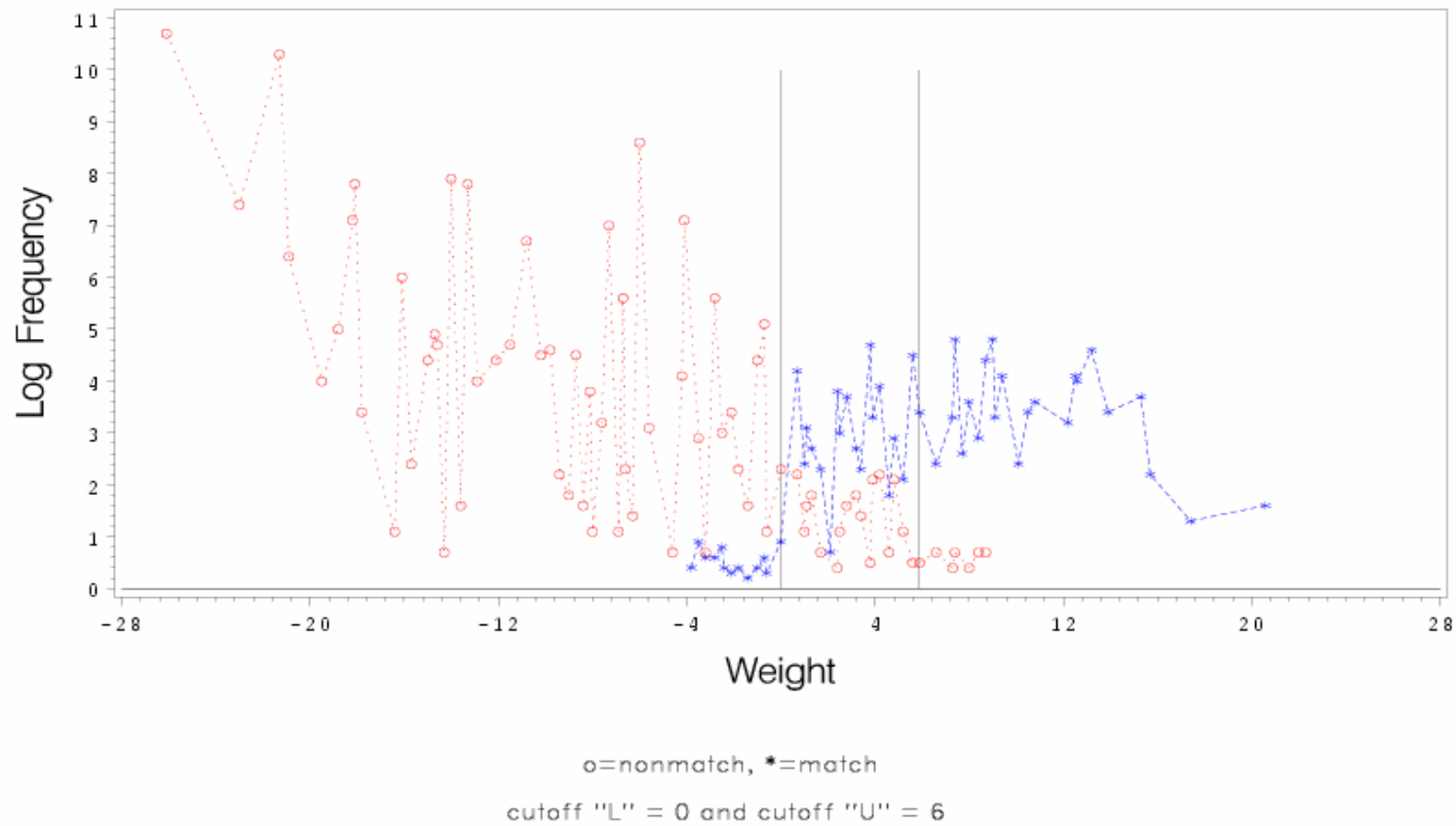
If  $R > UPPER$ , then designate pair as a link (match).

If  $LOWER \leq R \leq UPPER$ , then designate pair as a possible link and hold for clerical review. (2)

If  $R < LOWER$ , then designate pair as a nonlink (nonmatch).

Rule is optimal in the sense that it minimizes the in-between region (under fixed upper bounds on the error in the 1<sup>st</sup> and 3<sup>rd</sup> regions).

Figure 1. Log Frequency vs Weight  
Matches and Nonmatches Combined



Need to estimate:  $P(\gamma | M)$  and  $P(\gamma | U)$

In simple situations without training data:

Can use large population file (Newcombe 1959, 1962).

Can use EM algorithm (Winkler 1988)

Can use random agreement for  $P(\gamma | U)$  and experienced guesses or followup for  $P(\gamma | M)$  (Winkler and Thibaudeau 1991)

*Difficulties:* parameters vary across pairs of files, parameters sometimes vary across geographic regions or based how file created

## 7. Estimate error rates (false negatives and false positives)

With suitable training data, straightforward.

Can also estimate from certain large population files based on matching file against itself (Newcombe 1959, 1962; Fellegi and Sunter 1969).

Can estimate with small amounts of training data in some situations (Larsen and Rubin 2001, Winkler 2002).

Can estimate without training data in some situations (Winkler 2006).

Figure 1a. Good Matching Scenario

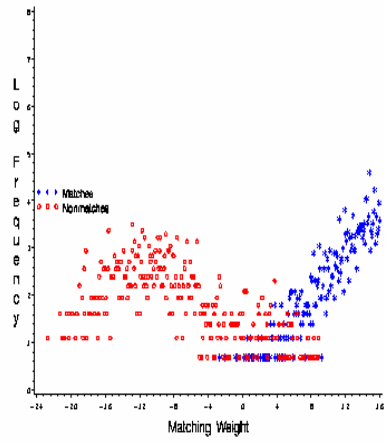


Figure 1b. Mediocre Matching Scenario

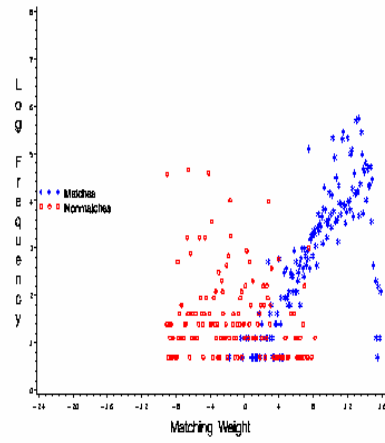


Figure 1c. 1st Poor Matching Scenario

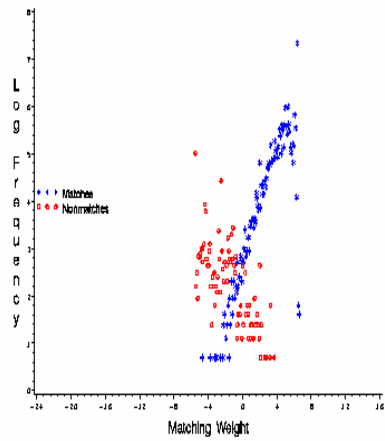
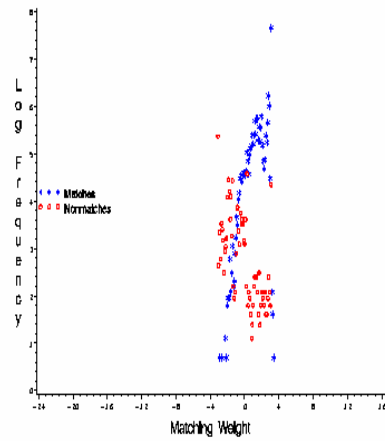


Figure 1d. 2nd Poor Matching Scenario



## 8. Determine effect of errors on estimation and analyses

*Regression* – Scheuren and Winkler (1993, 1997), Lahari and Larsen (2005)

*Loglinear models* – Winkler 1991

## Summarizing Issues

1. Need to have comparable information across files that provides distinguishing power.
2. May need file-specific methods of standardizing and parsing fields.
3. May need third file C to facilitate the matching of two files A and B.
4. Need estimates (bounds, crude guesses) of error rates.