

# Record Linkage and Machine Learning

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

<http://www.census.gov/srd/www/byyear.html>

March 11, 2004

Joint Program on Statistical Methodology

University of Maryland

## Outline

1. Introduction
2. Classical Record Linkage
3. Unsupervised Learning
4. Related Methods
5. Advanced Methods – Research
6. Concluding Remarks

# Examples of Record Linkage

## 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* – find duplicate records in health files using name, address, date-of-birth, etc. (Acheson - Oxford Record Linkage study)

Methods introduced by Newcombe (*Science* 1959, *CACM* 1962)

Divide product space  $A \times B$  of pairs from two files A and B into matches M and nonmatches U.

Consider *odds ratios*

$$R = P(\text{agree pat} \mid M) / P(\text{agree pat} \mid U)$$

agree pat = {agr/dis first, agr/dis last, agr/dis date-of-birth}

## Newcombe's Decision Rule

If  $R > \text{Upper}$ , call pair match;

If  $\text{Lower} \leq R \leq \text{Upper}$ , hold pair for clerical review;

If  $R < \text{Lower}$ , call pair a nonmatch.

Conditional Independence (Simplifying assumption)

$$\begin{aligned}\log R &= \log P(A_1, A_2, A_3 | M) / P(A_1, A_2, A_3 | U) \\ &= \log P(A_1 | M) / P(A_1 | U) + \log P(A_2 | M) / P(A_2 | U) + \\ &\quad \log P(A_3 | M) / P(A_3 | U)\end{aligned}$$

$A_i$  agree/disagree on  $i^{\text{th}}$  characteristic

Because  $U \sim A \times B$ , use  $P(A_i | U) \approx P(A_i)$ .

Take initial guess of  $P(A_i | M)$ .

After initial matching, take an (educated) set of pairs for review.  
Use a sample  $M_1$  of matches to approximate  
 $P(A_i | M) \approx P(A_i | M_1)$ .

Some similarity to Denis et al. (2003) ICML Workshop on Semi-Supervised Learning for text classification or type of batched, active learning.

Formal mathematical model introduced by Fellegi and Sunter (J. Amer. Stat. Assn. 1969). Rediscovered by Cooper and Maron 1978 JACM, others

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

$\gamma$  is an agreement pattern

If  $R > T_\mu$ , then designate pair as a match.

If  $T_\lambda \leq R \leq T_\mu$ , then designate pair as a potential match and hold for clerical review.

If  $R < T_\lambda$ , then designate pair as a nonmatch

$\mu$  - bound on false match rate

$\lambda$  - bound on false nonmatch rate.

Theorem FS (1969). Above decision rule is optimal in the sense that, for fixed bounds on the rate of false matches and nonmatches, it minimizes the clerical review region.

Figure 1. Log Frequency vs Weight Links

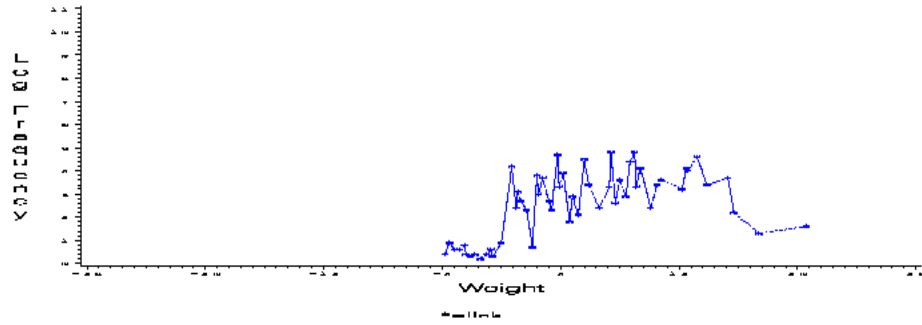


Figure 2. Log Frequency vs Weight Nonlinks

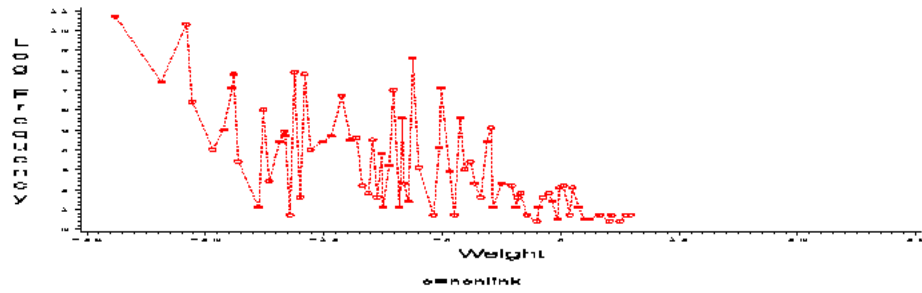
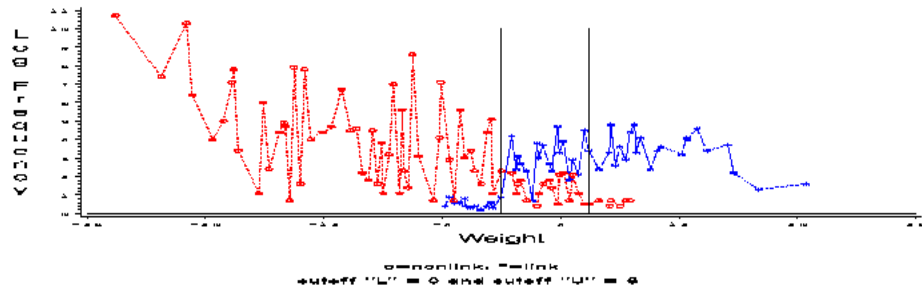


Figure 3. Log Frequency vs Weight Links and Nonlinks Combined



## No Training Data

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

Optimal parameters vary significantly from one region to the next in the 1990 U.S. Census (Winkler *ARC* 1989)

Software (Winkler and Thibaudeau 1991) finds optimal yes/no parameters automatically, builds frequency tables automatically that are scaled to yes/no parameters. Entire U.S. (450 regions in 1990) matched in three weeks.

$P(\text{agree first} \mid M)$ ,  $P(\text{agree last} \mid M)$  vary significantly  
Typographical error rates differ in adjacent regions

Do not need truth data set. Find optimal parameters (nearly automatically)

Fellegi-Sunter (FS) – 3 variables, independence

Winkler 1988 EM, independence

Winkler (1989a,b, 1993) general interaction accounting for dependence, convex constraints to predispose probabilities to appropriate regions, relative frequency (Smith vs Zabransky) (Della Pietra et al. 1997 *IEEE PAMI*, Winkler 1990 *Ann Prob*)

Larsen 1994, 1996 MCMC

Belin and Rubin JASA 1995 EM – error rates

Larsen and Rubin JASA 2001 MCMC

Some papers (e.g. Winkler rr94/05) available at

<http://www.census.gov/srd/www/byyear.html>.

<http://www.fcsm.gov/working-papers/wwinkler.pdf> overview

Elfekey, Vassilios, Elmagarmid IEEE ICDE 2002 –  
training and no training

Yancey ASA 2002 - no training – choose appropriate  
set of pairs

Winkler 2002 combinations of unlabelled and very  
small amounts of labeled training data  
(text classification, Nigam, McCallum, Thrun, and  
Mitchell, *Machine Learning* 2000, Winkler 2000)

Do and Rahm, VLDB '02 – training data

Sarawagi and Bhamidipaty VLDB '02 – training data

Bilenko and Mooney 2003, KDD '03 - training data

Give overall framework for evaluating methods

## Names (also can do addresses)

Starting with a free-form name, how do we get at the components that need to be compared? (Winkler software 1993 - agriculture)

Table Examples of Name Parsing

### Standardized

1. DR John J Smith MD
2. Smith DRY FRM
3. Smith & Son ENTP

### Parsed

	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith	Son		ENTP	

## Dealing with typographical error

*Bigram, Edit Distance* (classic computer science methods)

*Jaro and Winkler string comparators* (Winkler ASA 1990)

Winkler extensions related to ideas of Pollock and Zamora,  
1984 *Communications ACM*, modeling using test decks

Bigram easiest to compute, fastest – pairs of characters in common  
between two strings

Edit distance – slowest – dynamic programming, min number of  
insertions, deletions, substitutions to get from one string to another

In following, all string comparators scaled between 0.0 and 1.0,  
where 1.0 represents exact agreement

Table Proportional Agreement by String Comparator Values  
Among Matches, Key Fields by Geography

	StL	Col	Wash
First			
$\Phi=1.0$	0.75	0.82	0.75
$\Phi\geq 0.6$	0.93	0.94	0.93
Last			
$\Phi = 1.0$	0.85	0.88	0.86
$\Phi\geq 0.6$	0.95	0.96	0.96

$$\Phi_n (\text{Smith, Smith}) = 1.0$$

$$\Phi_n (\text{Dixon, Dickson}) = 0.8533.$$

Table Comparison of String Comparators Using  
Last Names, First Names, and Street Names

Two strings		String comparator Values			
		Jaro	Winkler	Bigram	Edit
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.925	0.818
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.917	0.889
NICHLESON	NICHULSON	0.926	0.956	0.906	0.889
JONES	JOHNSON	0.790	0.832	0.000	0.667
MASSEY	MASSIE	0.889	0.933	0.845	0.667
ABROMS	ABRAMS	0.889	0.922	0.906	0.833
HARDIN	MARTINEZ	0.000	0.000	0.000	0.143
ITMAN	SMITH	0.000	0.000	0.000	0.000

Table (cont) Comparison of String Comparators Using  
Last Names, First Names, and Street Names

Two strings		String comparator Values			
		Jaro	Winkler	Bigram	Edit
JERALDINE	GERALDINE	0.926	0.926	0.972	0.889
MARHTA	MARTHA	0.944	0.961	0.845	0.667
MICHELLE	MICHAEL	0.869	0.921	0.845	0.625
JULIES	JULIUS	0.889	0.933	0.906	0.833
TANYA	TONYA	0.867	0.880	0.883	0.800
DWAYNE	DUANE	0.822	0.840	0.000	0.500
SEAN	SUSAN	0.783	0.805	0.800	0.400
JON	JOHN	0.917	0.933	0.847	0.750
JON	JAN	0.000	0.000	0.000	0.667

## **Adaptive string comparators (Hidden Markov)**

W. Cohen et al. (IJCAI 2003, KDD 2003) – general string comparator

Yancey (ASA 2003) – typically census-type data

Bilenko & Mooney (KDD 2003) – entire free-form names, addr

Bilenko et al. (IEEE Intel. Sys. 2003)

## **Related work**

String Comparators – Ristad and Yanilios 1998 IEEE PAMI

Hidden Markov models, Viterbi algorithm

Wei 2004 IEEE PAMI – Markov Edit Distance

Address Standardization – Borkar, Deshmukh, Sarawagi 2001

ACM SIGMOD – Hidden Markov – very adaptive - mod Viterbi

Asian addresses

Churches, Christen, Zhu 2002, Christen, Churches, Lu, Zhu 2002

**Issues:** Estimate false match rates automatically without training data?  
Belin and Rubin (1995 JASA) - holds for easiest 5%

With training data, regression problem (Vapnik 2000, Hastie, Tibshirani, Friedman 2001) considered exceptionally difficult.

Small amounts of labeled combined with unlabeled (semi-supervised) -  
Larsen and Rubin (2001 JASA), Winkler (2002 ASA).

Cozman et al. 2003 ICML Workshop - Unlabeled data must have same model and come from same distribution as labeled data. Can these ideas be used to leverage small (or no training data)? Are there model classes that will work well (within  $\epsilon$ ) even though they are not exactly correct?

Estimation of false nonmatch rates without followup. Capture-recapture ideas of Sekar and Deming (1949) applied by Winkler (1989, also 1995).

## **Training and Test Data (Semi-supervised learning)**

1990 Census Data with truth and corrections

Only consider pairs agreeing on Census block and first character of last name

*Person:* First Name, Age, Marital Status, Sex

*Households:* Last Name, House Number, Street Name, Phone

	<b>Files</b>		<b>Files</b>		<b>Files</b>	
	<b>A<sub>1</sub></b>	<b>A<sub>2</sub></b>	<b>B<sub>1</sub></b>	<b>B<sub>2</sub></b>	<b>C<sub>1</sub></b>	<b>C<sub>2</sub></b>
<b>Size</b>	<b>15048</b>	<b>12072</b>	<b>5022</b>	<b>5212</b>	<b>4539</b>	<b>4851</b>
<b># pairs</b>	<b>116305</b>		<b>37327</b>		<b>38795</b>	
<b># matches</b>	<b>10096</b>		<b>3623</b>		<b>3490</b>	

Three classes of pairs

1. Matches within household
2. Nonmatches within household
3. Nonmatches outside household

Training Data Counts with Proportions of Matches

	A	B	C
Large Sample	7612 (.26)	3031 (.29)	3287 (.27)
Small Sample	588 (.33)	516 (.26)	540 (.24)

1-1 matching, non-1-1 matching

Large Sample approximately 8% of pairs

$$P(\gamma_i | \Theta) = \sum_j |C_j| P(\gamma_i | C_j; \Theta) P(C_j; \Theta) \quad (4)$$

$$P(\gamma_i | C_j; \Theta) = \prod_k P(\gamma_{i,k} | C_j; \Theta) \quad (5)$$

$$P(\Theta) = \prod_j (\Theta_{C_j})^{\alpha-1} \prod_k (\Theta_{\gamma_{i,k} | C_j})^{\alpha-1} \quad (6)$$

$$l_c(\Theta | D; z) = \log ( P(\Theta) ) + \\ (1-\lambda) \sum_{i \in D_u} \sum_j z_{ij} \log ( P(\gamma_i | C_j; \Theta) P(C_j; \Theta) ) + \\ \lambda \sum_{i \in D_l} \sum_j z_{ij} \log ( P(\gamma_i | C_j; \Theta) P(C_j; \Theta) ). \quad (7)$$

where  $0 \leq \lambda \leq 1$ .

2 or 3 Classes  $C_i$ , Equation (5) conditional independence,  
Equation (6) Dirichlet prior ( $\alpha < 1.1$ ),  
also general interaction (Winkler 1989, 1993, Larsen and Rubin 2001)

EM issues (3-class EM:  $C_1$ - match within household,  $C_2$  - nonmatch within household,  $C_3$  – nonmatch outside household)

1. Models – CI – independent –  $i1$  ( $i0$  – 1990 version)  $I,I,I$   
Larsen-Rubin CI in class 1, 4-way person,  
4-way household in classes 2 and 3,  $g1$   $I,HP,HP$   
Winkler 4+ way interactions in all classes,  $g3$  ( $g0$  1990 version)
2. lambda – how much to emphasize training data
3. delta – 0.000001 to 0.001 – smooth out peaks ( $\delta = \alpha - 1$ )
4. how many iterations (Friedman 2001, numerous ICML 2003)
5. number of degrees of partial agreement
  - a. agree, disagree (and/or blank) [small base = 2]
  - b. very close agree, moderately close agree, somewhat agree, blank, disagree [large base = 5]

metaparameters – Hastie, Tibshirani, Friedman 2001, Friedman 2001

## Measures of Success

Accuracy under 1-1 matching

Proportion of matches at given error levels

Accuracy under non-1-1 matching

Accuracy of estimates of error rates

(left tail matches, right tail nonmatches, overlap region)

(compare truth – 45 degree line against cumulative plot of estimates against truth)

Vapnik 1999

Hastie, Tibshirani, Friedman 2001

As parametric form of model has more parameters, need much more training data. Is it possible to leverage unlabeled data?

## 1990

yes/no

CI (i0), interact (g0)

I,I,I ; HP+,HP+,HP+

1-1, non-1-1

no delta

## 2002

3-level yes, blank, no

CI (i1), interaction (g1, g3)

I,I,I ; I,HP,HP; HP+,HP+,HP+

1-1, non-1-1

delta smoothing

First, Last, Hsnm, Stnm, Phone, Age, Marital, Sex

$5 \cdot 5 \cdot 5 \cdot 5 \cdot 5 \cdot 5 \cdot 3 \cdot 3 = 140625$  patterns – large base

With small base 2, there are 256 patterns

Work intended as exploratory/data-mining approach; determine methods that work best, leverage small amounts of training data

## **Overall**

1990 procedures good, very difficult to improve

Best of new procedures

Unsupervised learning for 1-1 matching

interactions +0.5%

additional partial agreements +0.5%

Mixtures of labeled and unlabeled for non-1-1 matching

Relatively accurate measures of error rates

**Note:** In record linkage applications, the quality of results is primarily dependent on the quality of the information in input files. 1990 Census data of relatively high quality for record linkage.

**Table1. Matching efficacy, 1-1 matching**

**Error level (accuracy)**

	File A	File B	File C
0.002			
g3	9780 (0.967)	3428 (0.944)	3225 (0.922)
g1	9741 (0.965)	3448 (0.950)	3261 (0.932)
i1	9640 (0.956)	3277 (0.903)	3042 (0.867)
i0	9701 (0.959)	3489 (0.961)	3306 (0.945)
g0	9649 (0.954)	3422 (0.943)	3273 (0.936)
0.005			
g3	9882 (0.974)	3547 (0.974)	3409 (0.972)
g1	9868 (0.973)	3523 (0.967)	3386 (0.965)
i1	9855 (0.971)	3513 (0.965)	3314 (0.945)
i0	9857 (0.971)	3540 (0.972)	3379 (0.963)
g0	9810 (0.967)	3511 (0.964)	3329 (0.949)

02:g3(HP,HP,HP),g1(I,HP,HP),i1(I,I,I); 90:g0=g3,i1=i0

## Table 1 (cont)

0.010

g3	9955	(0.976)	3584	(0.979)	3452	(0.979)
g1	9948	(0.976)	3568	(0.975)	3441	(0.976)
i1	9942	(0.975)	3566	(0.974)	3414	(0.968)
i0	9952	(0.976)	3580	(0.978)	3431	(0.973)
g0	9878	(0.969)	3536	(0.966)	3372	(0.956)

0.020

g3	10062	(0.976)	3622	(0.980)	3491	(0.980)
g1	10057	(0.976)	3614	(0.978)	3487	(0.979)
i1	9942	(0.976)	3614	(0.978)	3481	(0.977)
i0	10065	(0.977)	3623	(0.980)	3489	(0.980)
g0	9998	(0.970)	3589	(0.971)	3417	(0.960)

02:g3(HP,HP,HP),g1(I,HP,HP),i1(I,I,I); 90:g0=g3,i1=i0

Figure 7a. Estimates vs Truth, File A  
Cumulative Distribution of Matches

Estimated Large base, Independent EM, non-1-1

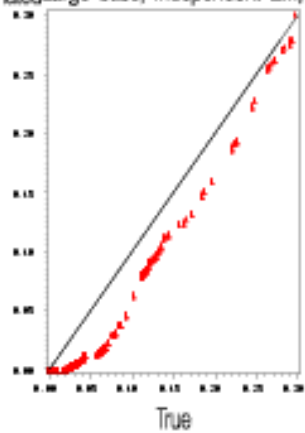


Figure 7c. Estimates vs Truth, File B  
Cumulative Distribution of Matches

Estimated Large base, Independent EM, non-1-1

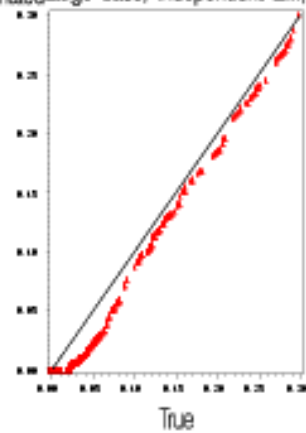


Figure 7e. Estimates vs Truth, File C  
Cumulative Distribution of Matches

Estimated Large base, Independent EM, non-1-1

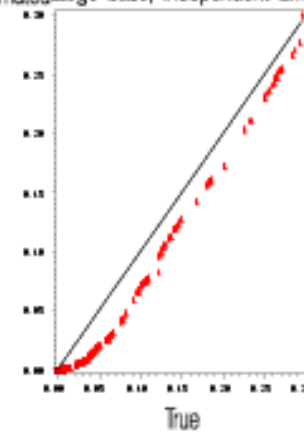


Figure 7b. Estimates vs Truth, File A  
Cumulative Distribution of Nonmatches

Estimated Large base, Independent EM, non-1-1

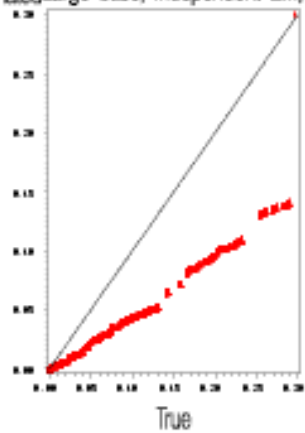


Figure 7d. Estimates vs Truth, File B  
Cumulative Distribution of Nonmatches

Estimated Large base, Independent EM, non-1-1

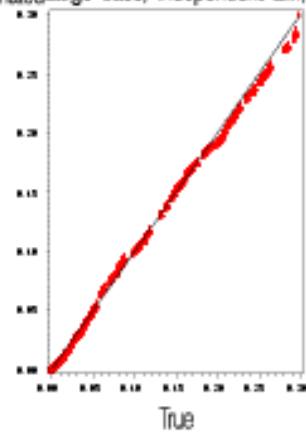


Figure 7f. Estimates vs Truth, File C  
Cumulative Distribution of Nonmatches

Estimated Large base, Independent EM, non-1-1

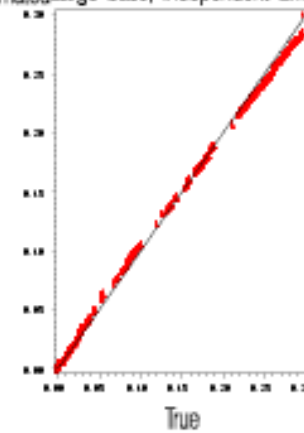


Figure 8a. Estimates vs Truth, File A  
Cumulative Matches, Lambda=0.9  
Small Sample, Independent EM, non-1-1

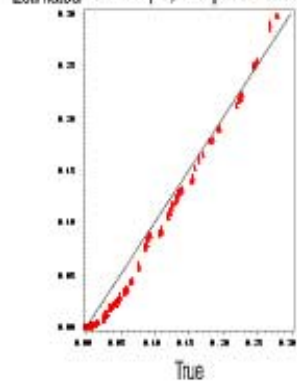


Figure 8c. Estimates vs Truth, File B  
Cumulative Matches, Lambda=0.9  
Small Sample, Independent EM, non-1-1

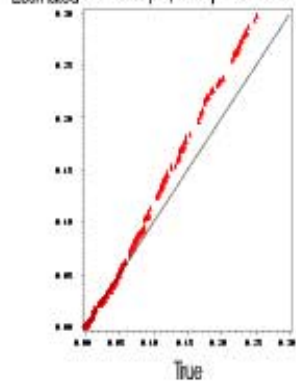


Figure 8e. Estimates vs Truth, File C  
Cumulative Matches, Lambda=0.9  
Small Sample, Independent EM, non-1-1

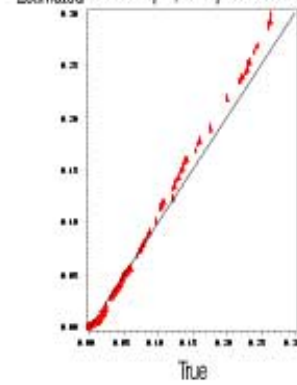


Figure 8b. Estimates vs Truth, File A  
Cumulative Nonmatches, Lambda=0.9  
Small Sample, Independent EM, non-1-1

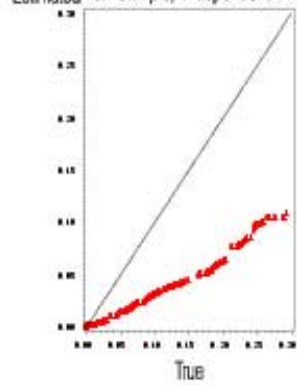


Figure 8d. Estimates vs Truth, File B  
Cumulative Nonmatches, Lambda=0.9  
Small Sample, Independent EM, non-1-1

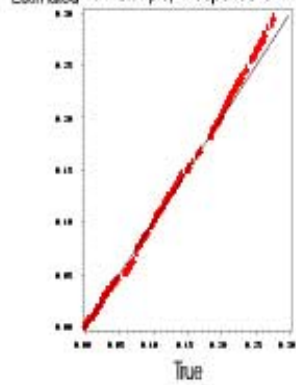
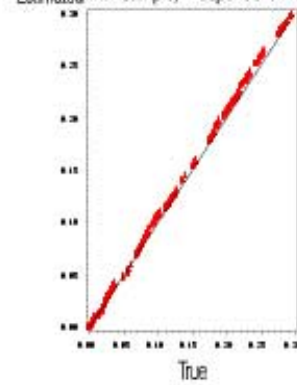


Figure 8f. Estimates vs Truth, File C  
Cumulative Nonmatches, Lambda=0.9  
Small Sample, Independent EM, non-1-1



## **BigMatch Software** (Yancey and Winkler 2003, 2004)

Match moderate size list having 100 million records against large administrative lists having upwards of 4 billion records  
5%+ typographical error in each field

10 sets of blocking criteria, no sorting of files  
New indexing, very fast retrieval and comparisons

Bentley and Sedgewick 1997 *ACM/SIAM Joint Conference on Discrete Algorithms*

SocSecAdmin – 600 million; Census – 300 million; Calif  
Quarterly Employment, 20 yrs – 1 billion

**Real-world Problem:** Hypothetical typographical error rates:

First Name – 0.20, Last Name – 0.10

Year-of-birth – 0.20, Month-of-birth – 0.20, Day-of-birth – 0.20

*Proportion of Duplicates Found* (using exact character-by-character matching on first name, last name, and full date-of-birth)

~0.10 – first name and last name combination are not commonly occurring

~0.08 – reduction for commonly occurring first-last name pairs and moderate limitation of search area

*with half the typographical error rate in each identifying field*

~0.55 – first name and last name combination are not commonly occurring

~0.46 – reduction for commonly occurring first-last name pairs and moderate limitation of search area

**Issue:** Reduce number of pairs brought together during matching  
(Partially) account for typographical error

Do multiple blocking passes, during each pass only consider pairs agreeing on blocking criteria

Block 1 – agree on block id (~50 or 70 households), first character surname

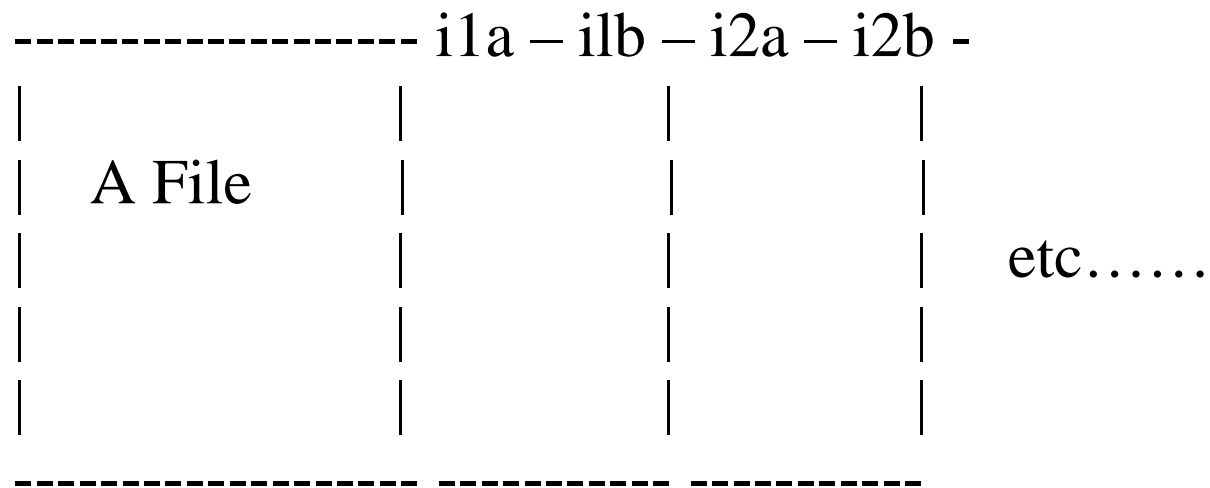
Block 2 – agree on block id, first character of street name

Block 3 – agree on surname, year-of-birth

Block 4 – agree on first character surname, first character first name, agree on US Postal ZIP code (~50000 individuals)

Developing a set of blocking criteria is very specific to the files being matched. It can also be dependent on the analyses on the linked files.

## Small File A and Large File B



IBM PC – 4-16 gigabytes RAM – 2GHz

FAST: 10 blocking passes, all sorts, all matches

100,000 pairs per second

US Census – 300 million records, 10 billion pairs, 3 days

Chaudhuri, Ganjam, Ganti, Motwanti 2003 ACM SIGMOD  
Characterize edit distance, approximate with q-grams, create indexes that are based on q-grams, use Chernoff bounds,  $10^3$  speed improvement over naive

Baxter, Christen, Churches, 2003 ACM SIGKDD Workshop  
Create q-gram indexes, 6 char word has 5 2-grams, create 5 indexes corresponding to agreeing on 4 of 5 2-qgrams

Jin, Li, Mehrota 2003 DAASFA '03

Li, Chang, Garcia-Molina, Wiederhold IEEE TKDE 2002  
Characterize distance (e.g. edit), imbed in  $R^d$ , StringMap, Combine R-trees

**Research:** Need characterization of typographical error among a set of fields and redundancy of fields. Gear search strategies to characterization.

Figure 1a. Good Matching Scenario

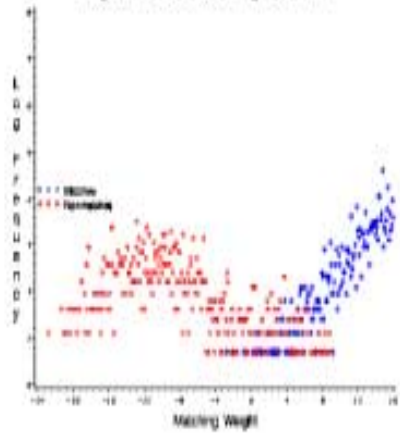


Figure 1b. Moderate Matching Scenario

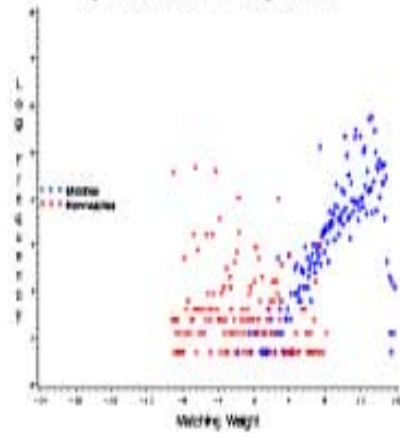


Figure 1c. 1st Poor Matching Scenario

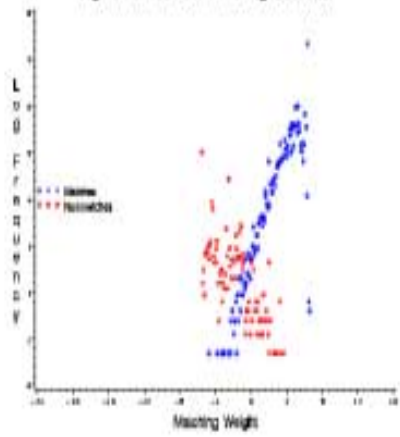
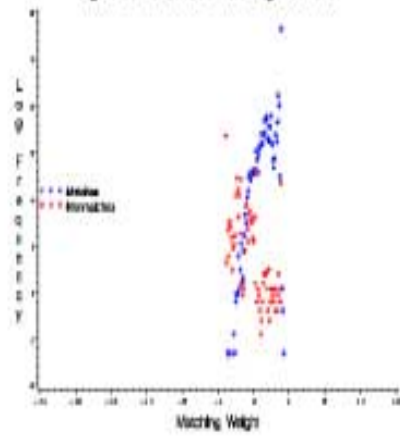


Figure 1d. 2nd Poor Matching Scenario



# No Unique Identifiers to Connect Records

## *Economics- Companies*

Agency A

Agency B

fuel ----->

outputs

feedstocks ----->

produced

## *Health- Individuals*

Receiving

Agencies

Social Benefits

B1, B2, B3

Incomes

Agency I

Use of Health

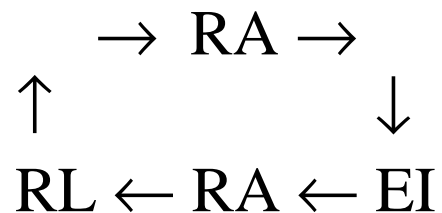
Agencies

Services

H1, H2

## Scheuren and Winkler (1997)

Analytic Linking takes the form



Do record linkage with name and address information only.

For moderate (200-500) number of pairs with high matching weight, do regression  $y = \beta x$  where  $y$  comes from one file and  $x$  comes from other file. Do outlier detection, build crude model, and put  $\text{pred}(y)$  in file having  $x$ . Perform matching with additional matching information. Iterate.

**File A**

**Common**

**File B**

**$A_{11}$  , ...  $A_{1n}$**

**Name1, Addr1**

**$B_{11}$  , ...  $B_{1m}$**

**$A_{21}$  , ...  $A_{2n}$**

**Name2, Addr2**

**$B_{21}$  , ...  $B_{2m}$**

**.**

**.**

**.**

**.**

**.**

**.**

**$A_{N1}$  , ...  $A_{Nn}$**

**NameN, AddrN**

**$B_{N1}$  , ...  $B_{Nm}$**

Using  $A_{ij}$  data, create a variable Pred ( $B_{kl}$ ) corresponding to B-File variables that is added to File A. Can iterate to improve Pred ( $B_{kl}$ ) and overall set of linkages.

Intuitive Idea: Name and Address may only allow 5-20% matching with business. Geographic identifiers and other variables may allow some clustering (i.e., record in A file can be associated with at most 60-500 records in B file). Quantitative information gives linkages.

Simpler situations: Scheuren and Winkler (1993), Lahiri and Larsen 2000, 2004.

More general situations: Scheuren and Winkler (1997), Larsen (2002), Winkler (2004, in progress)

## **Observations, Enhancements**

1. Small sample of (high-weight) pairs to get functional relationship between x-variables in A-file and y-variables in B-file. Alternatively, economists give relationships. Or have small amounts of training data.
2. Functional relationship defines metric(s) that can be used to improve matching (particularly among high weight pairs).
3. Use name, address (or other) information to weakly cluster record from A-file with small number of records in B-file.  
Use functional relationship to improve matching.
4. Iterate while improving functional relationships among highest weighted pairs. Gradually reduce (or remove) dependence on name and address.

Descending sort of variables.

$x_{i1}$

$x_{i2}$

•

$x_{ik}$

•

•

$x_{il}$

•

•

$x_{in}$

$y_{j1}$

$y_{j2}$

•

$y_{jq}$

•

•

$y_{jr}$

•

•

$y_{jm}$

Set of functions  $f_{ij}$  that relate  $x_i$  in file A to  $y_j$  in file B.

May only need two or three functions.

Based on highest weight pairs, partition ranges of x- and y-variables, x- and y-variables are weak identifiers.

Can build a function  $f_{ij}$  relates a range of  $x_i$  with a given probability, a second pair of ranges with a lower probability, and (possibly) a third range with a very low probability.

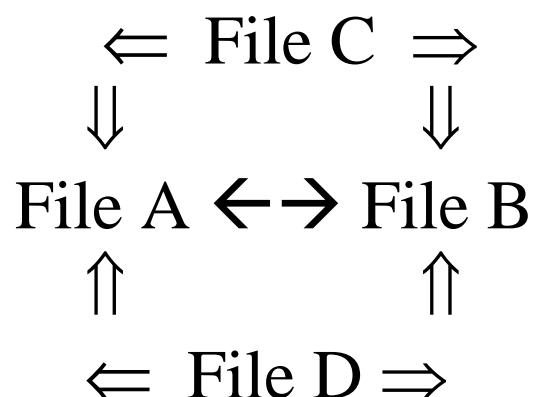
Each  $f_{ij}$  can be quite inaccurate. The redundancy of information from several  $f_{kl}$  may be sufficient for linkage.

*For individual:* mortgage  $\leftrightarrow$  house value  $\leftrightarrow$  income

*For company:* receipts  $\leftrightarrow$  income

## Valid Analytic Relationships $\leftrightarrow$ Metrics for Comparing Records in Files

Training data can give a straightforward way of getting functional relationships and metrics (e.g., Neural Nets).



Bridging File (Winkler, 1999)  
(extra variables – population file)

Reduce # pairs can be linked to

Identity Uncertainty – S. Russell, Proc. IFSA '01, Pasula et al.  
NIPS 03

Probabilistic Relational Models – (assume training data)

Koller and Pfeffer 1998 – Proc AAAI

Getoor, Friedman, Koller, Taskar 2001- ICML '01, JMLR '02

Lu and Getoor - ICML 2003 - linked based

Taskar, Wong, Koller - ICML 2003, also Taskar et al. UAI '02

Link Analysis – Lafferty, McCallum, Pereira KDD 2001,  
McCallum and Wellner IJCAI 2003, others

Optimally Weighted Aggregates – Torra 2000, 2001, 2003, also  
*Information Fusion in Data Mining*, Springer Studies in Fuzziness  
and Soft Computing, 2003.

$\mathbf{V}$  = set of discrete random variables

$G$  = undirected graph,  $C(G)$  = set of cliques in  $G$

$c \in C(G)$  associated with set of nodes  $\mathbf{V}_c$

clique potential  $\phi_c(\mathbf{V}_c)$

$\Phi = \{ \phi_c(\mathbf{V}_c) \}$

Markov net  $(G, \Phi)$  defines  $P(\mathbf{v}) = (1/Z) \prod_{c \in C(G)} \phi_c(\mathbf{V}_c)$  where product over  $\{ c \in C(G) \}$

$$\phi_c(\mathbf{V}_c) = \exp(w_c \cdot f_c(\mathbf{V}_c))$$

$$\log P(\mathbf{v}) = \sum_c w_c \cdot f_c(\mathbf{V}_c) - \log Z$$

$\mathbf{X}$  = set of random variables

$\mathbf{Y}$  = set of target (label) random variables

Conditional Markov network (random field)

$$P(\mathbf{y} | \mathbf{x}) = (1/Z(\mathbf{x})) \prod \phi_c (\mathbf{x}_c, \mathbf{y}_c) \quad (\mathbf{GenFrame})$$

$$\phi_c (\mathbf{x}_c, \mathbf{y}_c) = \exp (y w_k x_k )$$

Computation via belief propagation (Pearl 1988)

via graph partitioning (McCallum and Wellner 2003 IJCAI)

Relate every record (entity) to every other entity

$C(G)$  are given in many examples

Web pages – hyperlinks give typical links that must be weighted

Pairs in a record linkage situation

File  $\mathbf{A} = \{A_i : i = 1, \dots, n\}$ , File  $\mathbf{B} = \{B_j : j = 1, \dots, m\}$

In  $\mathbf{A} \times \mathbf{B}$ , every record is related to every other record

Blocking Pass (possibly)

Each  $A_i$  is associated with a subset of  $B_j$  s.

Discretize x-variables in  $A_i$  s and y-variables in  $B_j$  s

Choose subset of x- and y-variables to get functions  $f_{ij}$

Probabilities associated with  $f_{ij}$  give metrics, can iterate to refine probabilities, issue: in many situations, data are so good can do linkage manually (Malin, Sweeney, etc. CMU 2002-2004)

Need to extend (**GenFrame**)

Maybe need new framework for how to do modeling and how it related to data

Examples of how to do computation and get results in very simple situations (no training data, using economic relationships, distributional properties of x- and y-variables, information in auxiliary files)

Scheuren and Winkler (1997)

Winkler (2002) – microdata re-identification

Winkler (2004a, b) – synthetic microdata re-identification and models of information loss

Palley and Simonoff (1987) ACM TDBS