Discussion of “Entity Resolution: Measuring and Reporting Quality” by Rebecca Steorts

Goal: Combine sets of files to create larger, cleaner sets of data for policy analyses.

Economics- Companies

Agency A		Agency B
fuel	------>	outputs
feedstocks	------>	produced

Health- Individuals

Receiving Social Benefits	Agencies B1, B2, B3
Incomes	Agency I
Use of Health Services	Agencies H1, H2

File A	Common	File B
A_{11} , ... A_{1n}	Name1,Addr1,DOB1	B_{11},...B_{1m}
A_{21} , ... A_{2n}	Name2,Addr2,DOB2	B_{21},...B_{2m}
.		.
.		.
.		.
A_{N1} , ... A_{Nn}	NameN,AddrN,DOBN	B_{N1},...B_{Nm}

Issues:

1. Clean-up original source files (**A** and **B**)
    a. Modeling/edit/imputation
    b. Data linkage (duplication)
2. Create merged file (data linkage)
3. Adjust statistical analysis for linkage error
   (research problem, easiest 5-20% solved)
    a. Enhancements to current elementary models
    b. Extensions using modeling/edit/imputation and statistical matching

For 100s of millions of records, computational algorithms need to be 2-6 orders faster than those used previously.

Cautions

5% error in each of files **A** and **B**

5% matching error

Errors are additive

15% error in $(A_{j1}, \ldots A_{jn}, B_{j1}, \ldots B_{jm})$ data

Are there any analyses that are possible?

If the error is reduced to 5% overall, what analyses are possible?

How will we even know how much error is in the files?

Figure 1. Log Frequency vs Weight
Links

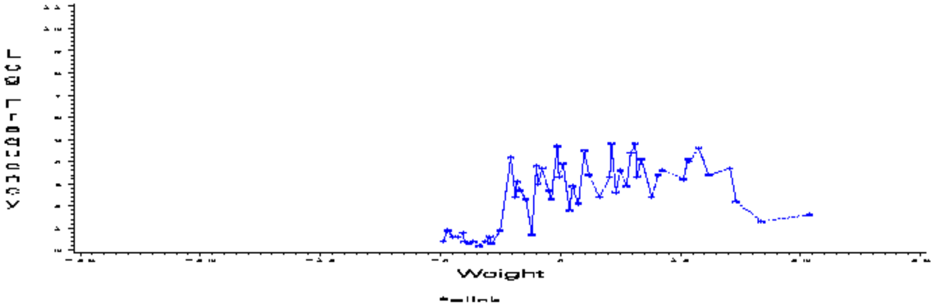


Figure 2. Log Frequency vs Weight
Nonlinks

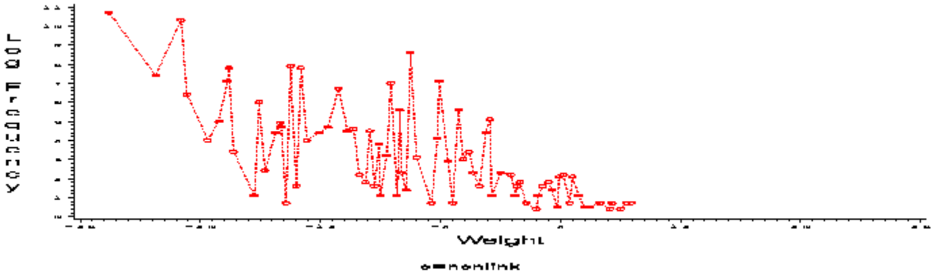
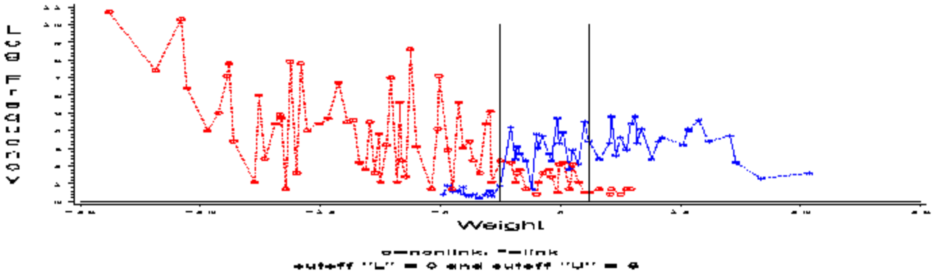


Figure 3. Log Frequency vs Weight
Links and Nonlinks Combined



Steorts points out that transitivity typically does not hold.

1 <-> 2, 2 <-> 3 does not necessarily yield 1 <-> 3.

Issue with both Fellegi-Sunter model and with the data.

Most systems still use FS model because the deviations may not be ‘too bad’.

An issue: In most record linkage situations, *there is never training data*.

Example from 1990 Decennial Census.

Needed to find ‘optimal’ parameters in ~500 regions without training data where parameters varied significantly from region to region. Matching needed to be completed in 3-6 weeks to meet production schedules. ‘Optimal’ parameters reduced clerical review region by 2/3 (only 200 individuals instead of 600 – Winkler 1989).

Unsupervised methods (Winkler 1988, also 2006) yield ‘optimal’ parameters. The methods outperform active learning (semi-supervised) where ‘truth’ of set of pairs is determined, the set of pairs being reviewed, the procedure is repeated until parameters converge. The methods were rediscovered by K. Larsen (2005 SIGKDD Explorations).

The 1990 clerical review region consisted almost entirely of individuals within in the same household who were missing both first name and age. In suburban regions, 1/40 of these ‘blanks’ were true matches; in urban regions 1/10 were true matches. These ‘blanks’ that were converted to true matches were 2% of the matches that were found. There were an additional small proportion (1-2%) that were estimated by the method of Belin (1990) and additional 0.5% of pairs converted to true matches via field follow-up where the pairs had no 3-grams in common.

The methods of Steorts (and a few others) are exceptionally promising and should improve Census methods and others' methods. Researchers need much better test decks (such as a few files available in the Decennial Census) for more careful evaluation.