william.e.winkler@census.gov

Table 1. Basic match situation

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}

Three Issues

1. Clean up individual **A** and **B** files
2. Improve record linkage using non-unique quasi-identifiers
3. Adjust analyses on linked files for linkage error

Issue 1: Modeling/edit/imputation according to Fellegi-Holt model (1976)
Winkler (2003, 2008, 2010)
100 million+ records – need algorithms ~100 times as fast as commercial

Issue 2. Improve Record Linkage according to Fellegi-Sunter model (1969)
No Training Data
Better parameter estimation : Winkler (1987, 1993), Ravikumar and Cohen (2004), Bhattacharya and Getoor (2006)
Estimate false match rates: Belin and Rubin (1995), Winkler (2006)
Estimate false nonmatch rates: Winkler (2004)
10^8-10^9 records: BigMatch is 40+ times as recent parallel software

Issue 3. Adjust Analyses for Linkage Error
Scheuren and Winkler (1993, 1997), Lahiri and Larsen (2005), Chambers (2009)

Bhattacharya and Getoor (2006)
Somewhat related work with MCMC – Larsen and Rubin (JASA 2001), Tancredi and Liseo (AoAS 2011)

Resolve references using multiple authors. Provide methods that improve over methods typically used at the Census Bureau.

Analogy: use multiple addresses (or telephone numbers) associated with individuals or multiple individuals within households

Issues: (1) models/parameters associated with typographical error and different representations of names, addresses, etc.
(2) scaling to situations that are 10^5+ times as large
(3) improving models/representations for subtle additional situations

Winglee, Valliant, and Scheuren (2005)

Basic method should be straightforward to develop and apply.
Determine whether method can be applied to a second data set.

$$w_r = log_2 \frac{\prod_{v=1}^{V} \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}}}{\prod_{v=1}^{V} \prod_{i=1}^{c_v} u_{vi}^{y_{rvi}}}$$

$$m_{vi}^{y_{rvi}}$$

Cautions: Belin (~1989) "New Approaches on Matching for Census Undercount", Belin (~1990) "Recent Developments in Calibrating Error Rates for Computer Matching"
Alternative: apply Belin-Rubin or some other method.

Dong, Berti-Equille, and Srivastava (2009)
Many record linkage examples classifying articles, journals, authors
As use more files, have false information that can be reproduced