Module X: fastlink

Rebecca C. Steorts

Reading

- Binette and Steorts (2020)
- Others ??

Probabilistic Entity Resolution

While Fellegi and Sunter (1969) have provided a framework for probabilistic entity resolution, there are few implmentations that scale to large size data sets.

fastlink

- ▶ Edmorando et al. (2020) developed fastlink a scalable implementation of the FS method.

- ▶ In addition, the authors incorporated auxiliary information such as population name frequency and migration rates.

- ▶ The authors used parallelization and hashing to merge millions of records in a near real-time on a laptop computer, and provided open-source software of their proposed methodology.

Agreement Patterns

- Two data sets (A and B) with variables in common
- Agreement value in field a for record pair (i, j)

$$\rho_a(i,j) = \begin{cases} \texttt{agree} \\ \\ \texttt{disagree} \end{cases}$$

Agreement Patterns

- Two data sets (A and B) with variables in common
- Agreement value in field a for record pair (i, j)

$$\rho_a(i,j) \;=\; \begin{cases} \text{agree} \\ \\ \text{disagree} \end{cases}$$

	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	**Devereux St.**
Data set $\mathcal{B}$				
7	James	Smit	43	**Dvereux St.**
	agree	agree	disagree	**agree**