fabl: Fast Beta Linkage

Brian Kundinger

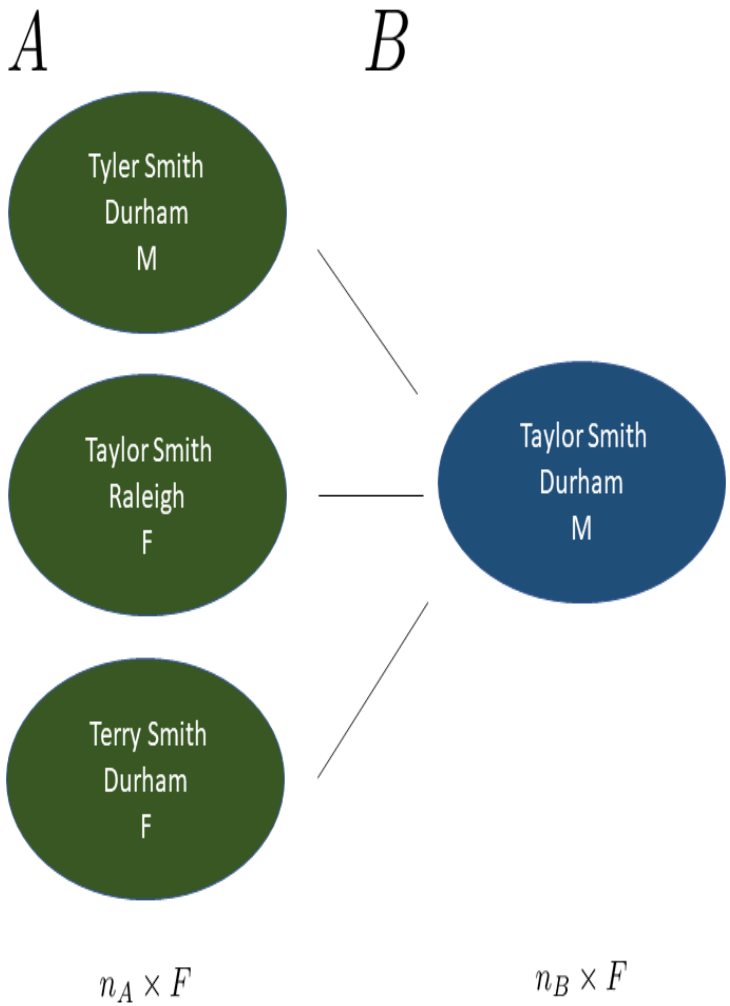
El Salvadoran Civil War

- El Salvadoran Civil War (1979 - 1992)
- Multiple organizations have documented casualties. El Rescate - Tutela Legal (ERTL) contains $n_A = 4420$ records, and Salvadoran Human Rights Commission (CDHES) contains $n_B = 1323$ records.
- We need to find duplicates across separate datafiles to accurately estimate *documented identifiable deaths* (DID)
- Not much information, only first name, last name, and date and place of death
- Spanish language: very common for parent and child to share the same name, "second" names often omitted

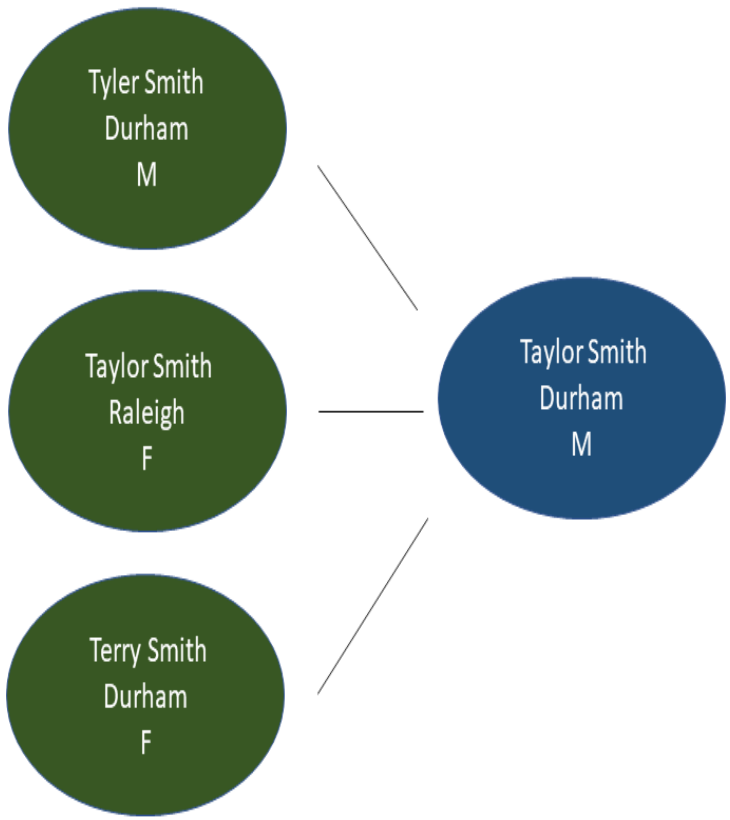
Foundations of Method

- Many record linkage methods are derived from Fellegi and Sunter's "A Theory for Record Linkage" (1969)
- Transform the F fields of information among the n_A records in File A and the n_B records in File B into a $n_A n_B \times F$ matrix of comparison vectors Γ.
- Comparisons for each field f are binned to integers $\{1, \ldots, L_f\}$, with 1 being the highest level of agreement
  - Common to use $\{1, 2\}$ to reflect exact agreement or disagreement, and $\{1, 2, 3\}$ to allow for partial agreement

Comparison Vector Example



Comparison Vector Example



$$\Gamma$$

$$\gamma_{11} = (2, 1, 1, 1)$$

$$\gamma_{21} = (2, 1, 2, 2)$$

$$\gamma_{31} = (3, 1, 1, 2)$$

$n_A \times F$ $n_B \times F$ $n_A n_B \times F$

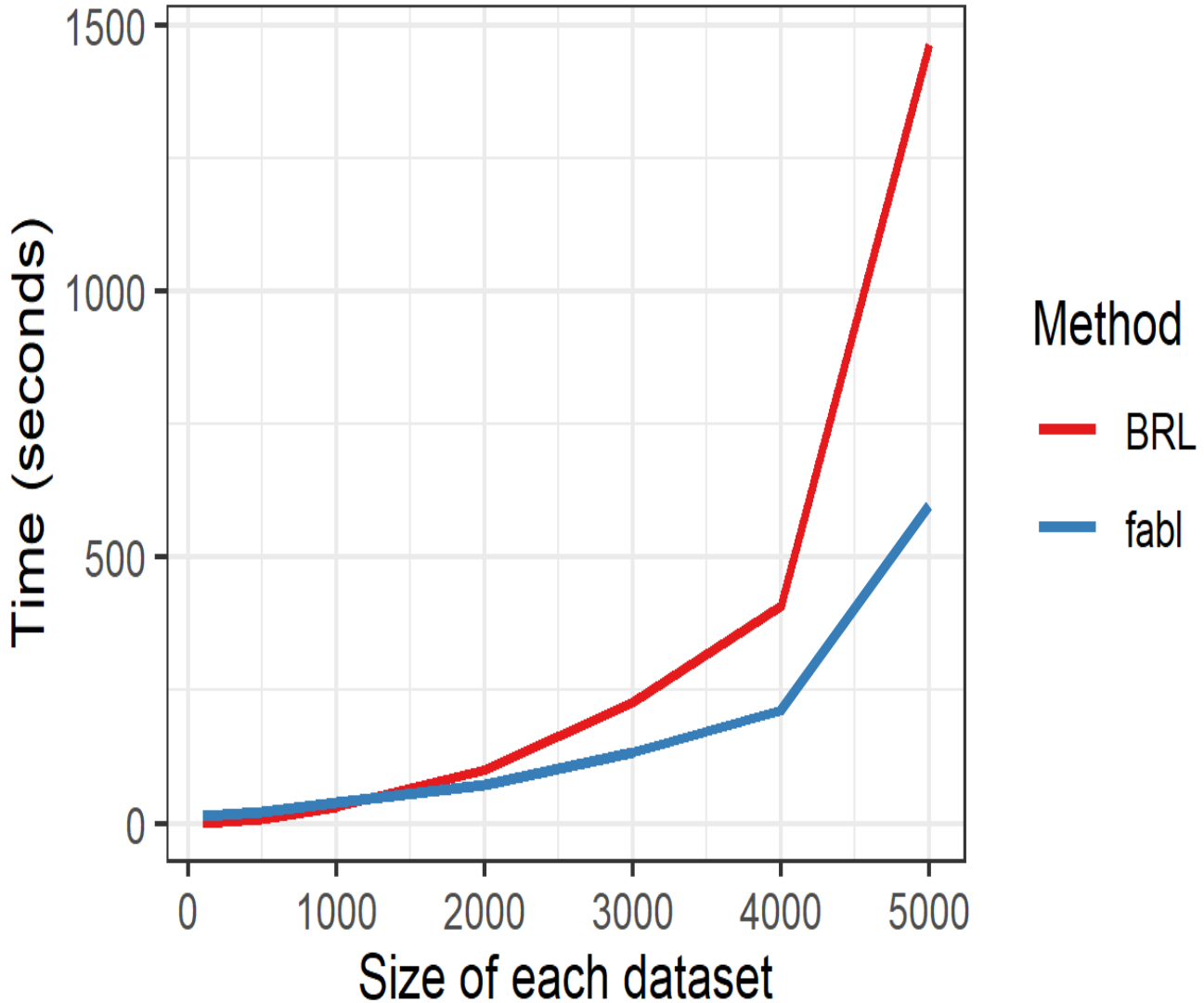
Previous Work

- Fellegi-Sunter (1969) made matching decisions on all $n_A n_B$ pairs independently. Fast, but can lead to severe overmatching.
- Sadinle (2017) proposed *Beta Record Linkage* (`BRL`), a Bayesian method that strictly enforces one-to-one matching. Matching proposals are *dependent*, resulting in a slow Gibbs sampler.

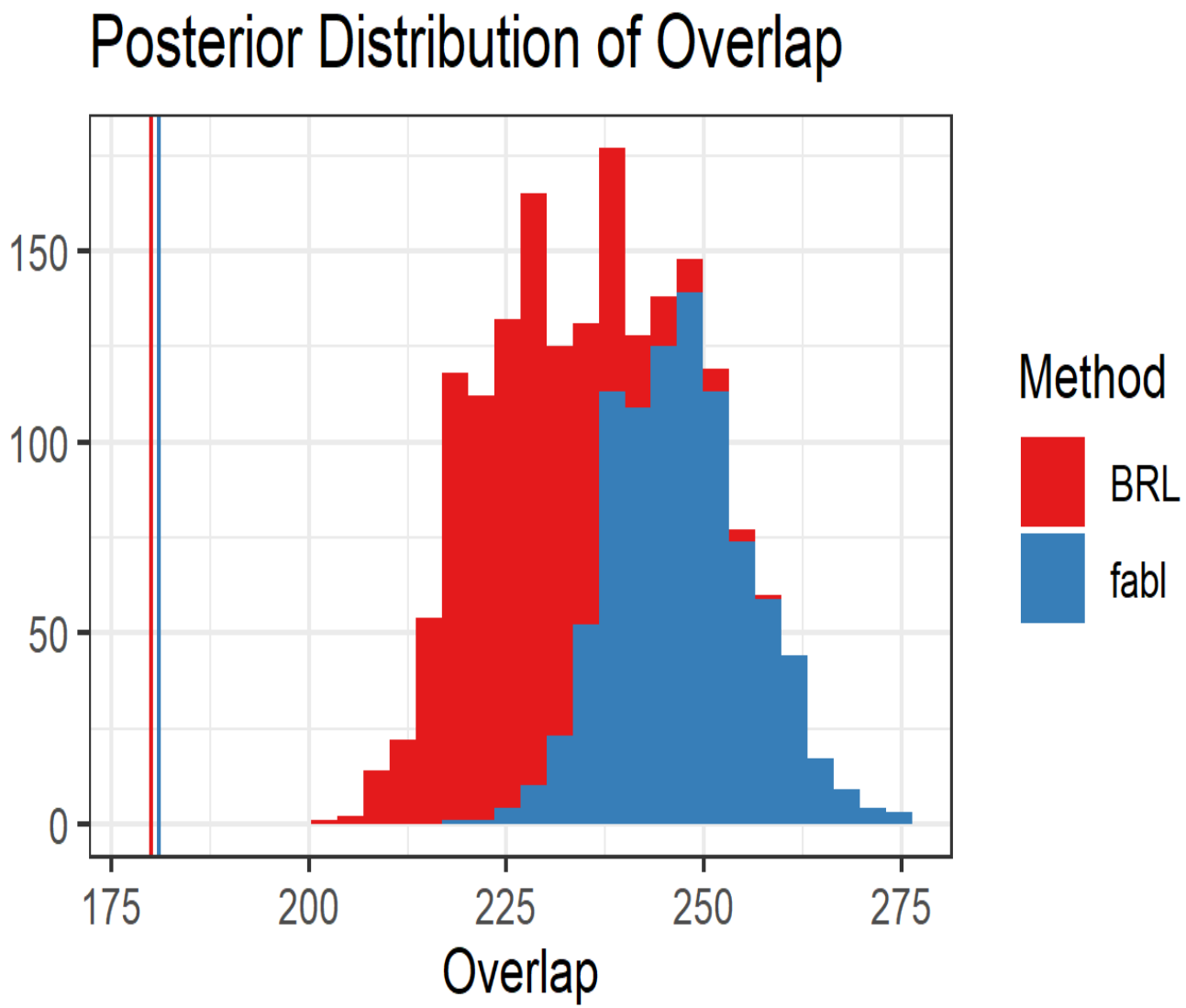
Contribution

- *Fast Beta Linkage* (`fabl`, pronounced "fable") makes one proposal per record in B, but makes these proposals *independently*. Allows for much quicker Gibbs sampler
- Hash comparison vectors to unique patterns, perform calculations over P patterns rather than $n_A n_B$ record pairs
- *Storage efficient indexing* to greatly use storage requirements of comparison vectors, allowing for larger linkage tasks.
- Gibbs sampler under `fabl` has computational complexity $O(n_B \times P \times F)$, significant savings compared to `BRL`'s $O(n_B \times n_A \times F)$

Simulation


1500
1000
500
0
Time (seconds)
0
1000
2000
3000
4000
5000
Size of each dataset
Method
BRL
fabl

El Salvador Data



El Salvador Data