Appendix

7.1 Summary of Fast Beta Linkage Method

procedure HASHING AND PREPROCESSING
  Construct and enumerate set of unique patterns $\mathcal{P}$ from F and $\{L_f\}$.
  Partition files $\boldsymbol{X_1}$ and $\boldsymbol{X_1}$ into chunks $\{\boldsymbol{X_{1n}}\}$, $\{\boldsymbol{X_{2m}}\}$.
  for each n, m **do**
    Create comparison vectors between $\boldsymbol{X_{1n}}$ and $\boldsymbol{X_{2m}}$.
    Hash records to $\mathcal{R}_{nm}$ and calculate summary statistics $\mathcal{H}_{nm}$.
    Use SEI to reduce memory usage; $\mathcal{R}_{nm} \rightarrow \mathcal{R}_{nm}^{\text{SEI}}$.
  end for
  Synthesize results across pairings to get $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{H}\}$.
end procedure
procedure GIBBS SAMPLING
  Initialize m, u, and Z parameters.
  for $t \in \{1, \ldots, T\}$ **do**
    Sample $\Phi^{t+1}|Z^t, \tilde{\Gamma}$.
    Sample $h\left(Z^{t+1}\right)|\Phi^{t+1}, \tilde{\Gamma}$. ▷ Sample agreement pattern, not record
  end for
  Sample $Z|h(Z), \tilde{\Gamma}$. ▷ Fills in record label based on agreement pattern
end procedure

7.2 One Hot Encoding Transformation

`fabl` makes use of one-hot encodings to aid in vectorized computations. For γ_{ij}^f with L_f levels, define e_{ij}^f to be an $L_f \times 1$ vector. When $\gamma_{ij}^f = \ell$, we set the ℓ^{th} element of e_{ij}^f to be 1, and set the other $L_f - 1$ elements of e_{ij}^f to be 0. We then concatenate the e_{ij}^f for all $f \in \{1, \ldots, F\}$, resulting in the one-hot encoded comparison vector e_{ij} of length $\sum_{f=1}^{F} L_f$.

For example, consider comparing the toy records shown in Table 6 with $L = (3, 3, 2, 2)$ levels of agreement for last name, first name, DOB, and city respectively. Since the first name differs by only one letter, a reasonable comparison vector for this pair would be $\gamma_{ij} = (1, 2, 1, 2)$. The one hot encoding representation of this vector is $e_{ij} = (1, 0, 0, 0, 1, 0, 1, 0, 0, 1)$.

Last Name	First Name	DOB	City
Smith	Taylor	01/01/2000	Durham
Smith	Tayler	01/01/2000	Raleigh

Table 6: Example records for one hot encoding.

7.3 Traceplots for Simulation Study

Below are traceplots for one of the 900 linkage tasks that comprise the simulation in Section 5.2. It is set up with one error across the linkage fields and 50 duplicates across files. Traceplots across other settings exhibit similar behavior. Note that traceplots for $\mathbf{u}$ parameters show very little variation because the overwhelming majority of record pairs are nonmatching.

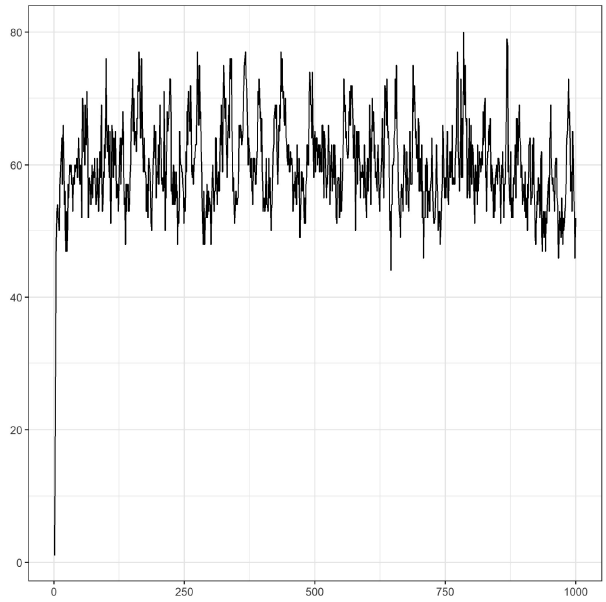

Figure 8: Representative traceplot of overlap between files from simulation study in Section 5.2

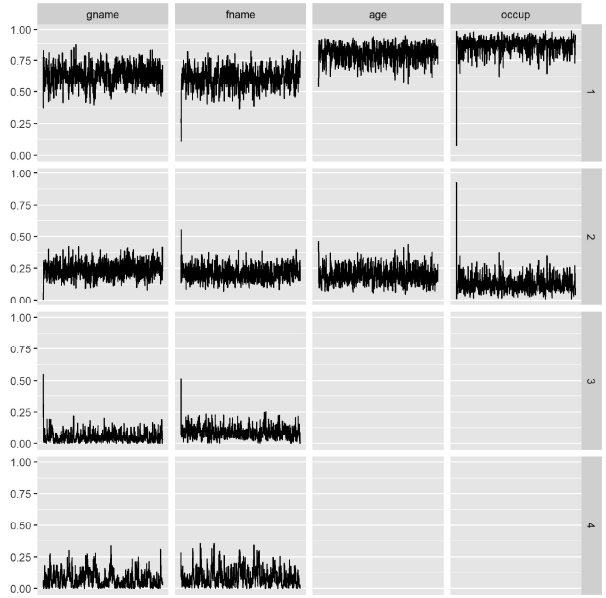

Figure 9: Representative traceplot of m parameter from simulation study in Section 5.2

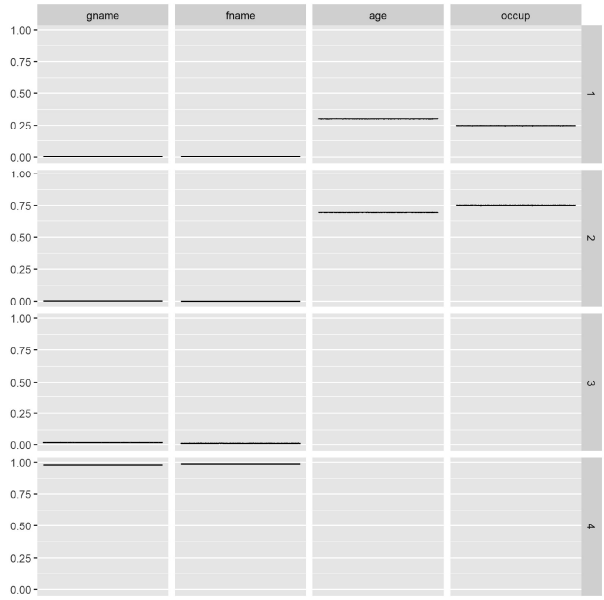

Figure 10: Representative traceplot of u parameters from simulation study in Section 5.2

7.4 Traceplots for El Salvador Case Study

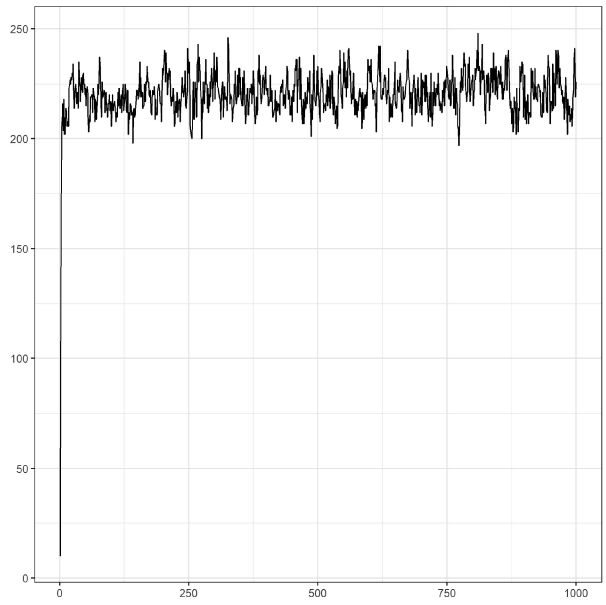

Figure 11: Traceplot for number of matches found across datasets in El Salvador case study

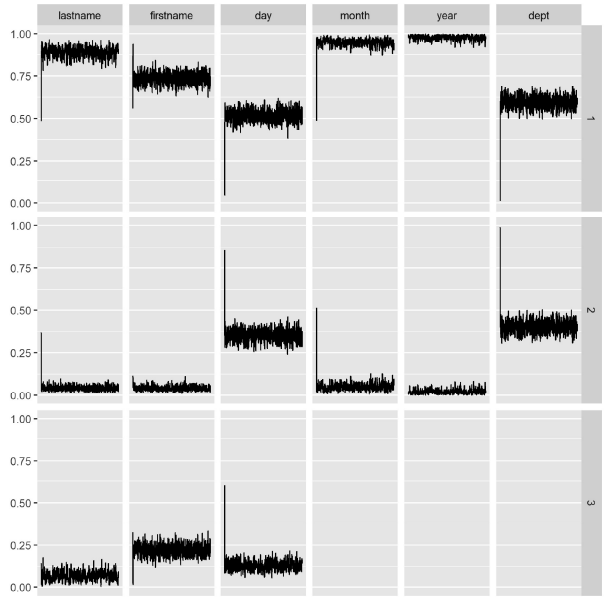

Figure 12: Traceplot for m parameter in El Salvador case study

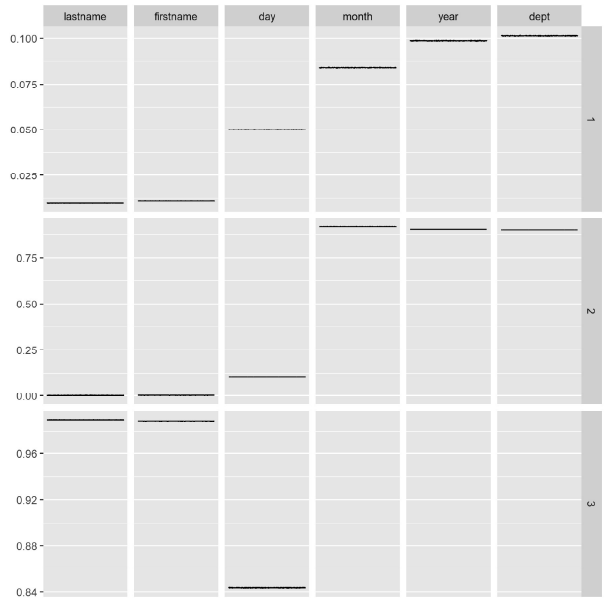

Figure 13: Traceplot for u parameter in El Salvador case study