A Genealogical Application of Statistical Record Linkage for Black Americans in the Antebellum South

Hannah Butler and Andee Kaplan
Department of Statistics, Colorado State University

Overview

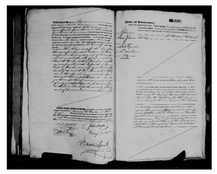
- **Data on individuals enslaved in the U.S. contain high rates of "alias" records which provide contextual or potentially overlooked information.**
- **We propose a statistical record linkage model that leverages aliases to link entities with greater success than existing models.**
- **The model is applied to digitized datasets to attempt to reconstruct more complete histories of these individuals.**

Data

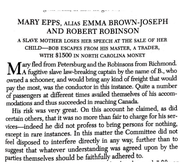
The form that data on enslaved African Americans in the antebellum south took was largely influenced by **oppression**.

- **Fugitive Advertisements**

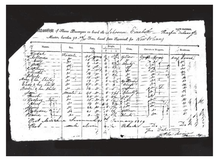
- **Notarial Records**



- **Narrative Accounts**



- **Ship Manifests**



Goal: Develop an effective statistical record linkage model for use in African American genealogical research through the antebellum era.

Alias Records

Alias: The occurrence of one or more duplications of an entity within a file, not due to error, but rather due to a known alternative piece of information.

Entity	Handle	Name	Account	Business	
User 1	@colorwired	Abby Joy	Personal	NA	Multiple pieces of valuable info
	@mysky.ceramics	My Sky Ceramics	Business	Artist	
User 2	@photogeanic	Emma Jean	Personal	NA	misleading/incorrect information
	@leafy_jean	E. Jean	Business	Grocery Store	
User 3	@oscarbutlerphoto	Oscar	Personal	NA	Reinforces record information
	@oscarbutlermusic	Oscar Butler	Personal	NA	
	@gettingfitwithoscar	Oscar Butler	Personal	NA	

What is Record Linkage?

Author	Title	Publ. Date
Stephen Jay Gould	The Mismeasure of Man	1996
Ian Hacking	The Taming of Chance	1990
Ruha Benjamin	Race After Technology	2019
Theodore Porter	Trust in Numbers	2020

Author	Title	Publ. Date
Meredith Broussard	More than a Glitch: Confronting Race, Gender, and Ability in Tech	2024
Ian Hackng	Taming of Chance	1990
Theodore M. Porter	The Rise of Statistical Thinking	2020
Michael E. Staub	The Mismeasure of Minds	2018

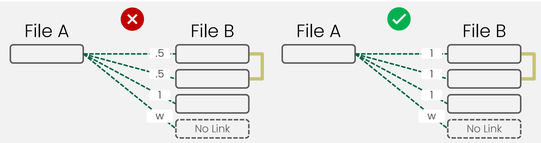
Record linkage encompasses a set of methods used to identify records belonging to a common entity.

- **Deterministic record linkage** uses rules to match field values. A unique identifier, such as an ISBN, makes this a straightforward task.
- **Statistical record linkage** is a statistical model used to estimate matches when no unique identifier is available. This approach offers more flexibility when data is error-prone, and often allows for uncertainty quantification.

Methods

- Specify a **Bayesian bipartite record linkage model** (BRLwA) that accommodates the presence of alias records.

- Choose a **linkage prior** to increase the prior probability that entities with multiple aliases have a greater probability of being linked.



Full Model

$$\mathbf{\Gamma} \mid \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{Z} \sim \prod_{i=1}^{N_1}\prod_{j=1}^{N_2}\prod_{f=1}^{F}\prod_{k=1}^{K_f}\left[(m_f(k))^{\mathbf{1}(Z_i=j)}(u_f(k))^{\mathbf{1}(Z_i\neq j)}\right]^{\gamma_{ij}^f(k)} \boldsymbol{V}_{ij}$$

$$\boldsymbol{m}_f \mid \boldsymbol{\alpha}_f \sim \text{Dirichlet}(\boldsymbol{\alpha}_f) \qquad \boldsymbol{u}_f \mid \boldsymbol{\beta}_f \sim \text{Dirichlet}(\boldsymbol{\beta}_f)$$

$$P(\boldsymbol{Z} \mid a, b) \propto \frac{\text{B}\left(a + L(\boldsymbol{Z}), b + N_1 - L(\boldsymbol{Z})\right)(N_2 - L(\boldsymbol{Z}))!}{\text{B}(a,b)N_2!}\mathbf{1}\left(\boldsymbol{Z}_i \neq \boldsymbol{Z}_{i'} \forall i \neq i'\right)$$

$\mathbf{\Gamma}$	Comparison data
$\boldsymbol{m}_f$	Agreement probabilities (linked)
$\boldsymbol{u}_f$	Agreement probabilities (unlinked)
$\boldsymbol{Z}$	Linkage vector

- Fit model using **MCMC** to generate posterior samples of the linkage vector and agreement probabilities.
- Use the posterior samples to **estimate links** across data files.

```
1. Initialize agreement probabilities and links

2. For each iteration, t = 1, . . ., T:
   1. Sample links, given data and current agreement
      probabilities
   2. Sample agreement probabilities, given data and current
      links

3. Use samples from above to estimate Z
```

Application

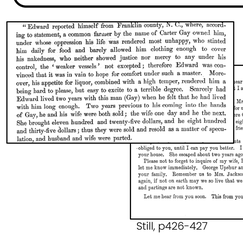
Fugitive Ads X Narrative Accounts

- **P(Alias) ~ 0.23;** 1,076 of 4,584 entities.
- estimated **~ 0.1% overlap;** 5 links by BLRwA.

- **Validation not available:**
  - 1 link manually confirmed.

Edward Lewis, alias William Brady. Estimated born in 1821. Enslaved in Franklin County, NC by Carter Gay. Attempted to escape enslavement in 1857.

Edgar. Estimated born in 1822. Enslaved by C H Gay in Franklin County, NC. Runaway slave advertisement published in 1857.


Still, p426-427


Freedom on the Move,

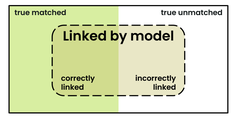
To-Do: Ship Manifests X Notarial Records

- **P(Alias) ~ 0.68;** 24,187 of 35,442 entities.

- **Validation available:**
  - **0.16% overlap;** 58 known matches.

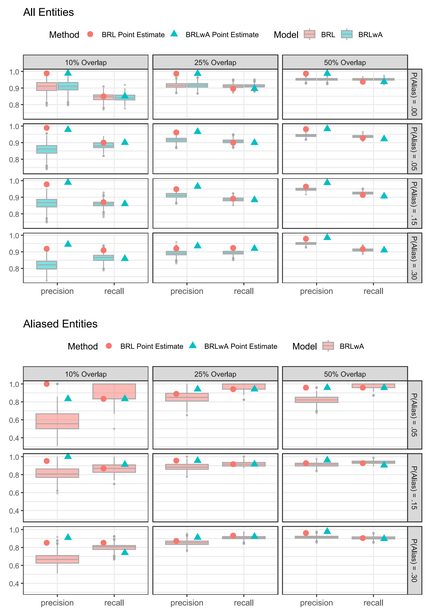
Simulation Results

- Tested BRLwA on simulated data against Bayesian model (BRL), with post-hoc conflict resolution.
- BRLwA point estimates of $\mathbf{Z}$ show comparable precision & recall to BRL (with correction).
- BRLwA maintains posterior distribution, allowing for uncertainty quantification.



Measuring Performance

Precision: *(correctly linked) / (all linked)*
Recall: *(correctly linked) / (true matched)*



References

- Fellegi, Ivan P., and Alan B. Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64, no. 328 (1969): 1183–1210. https://doi.org/10.2307/2286061.
- Freedom on the Move. "Freedom on the Move." Accessed October 30, 2024. https://freedomonthemove.org/.
- Nick Sacco. "William Still Underground Railroad Data.Xlsx." Accessed November 5, 2024. https://docs.google.com/spreadsheets/d/e/2PACX-1vTpXEcS1rao3f-ZHqeOApPoBXju6-Xeg902ujGjuDqkE_cxd0MGl8P-DFv4SMPgbg/pubhtml?usp=embed_facebook.
- Kinfolkology. "Kinfolkology." Accessed October 30, 2024. https://www.kinfolkology.org.
- Sadinle, Mauricio. "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association* 112, no. 518 (April 3, 2017): 600–612. https://doi.org/10.1080/01621459.2016.1148612.
- Still, William. *The Underground Railroad: A Record of Facts, Authentic Narratives, Letters &c., Narrating the Hardships, Hair-Breadth Escapes and Death Struggles of the Slaves in Their Efforts for Freedom*. Rev. ed. Philadelphia, Pa., Cincinnati, Ohio [etc.]: People's publishing company, 1879. https://www.loc.gov/resource/rbc0001.2019gen24984/.
- Still, William. "The Underground Railroad, Rev. Ed." Text. https://www.gutenberg.org/files/15263/15263-h/15263-h.htm. Accessed November 13, 2024. https://www.gutenberg.org/cache/epub/15263/pg15263-images.html.
- Williams, Jennie K. "Trouble the Water: The Baltimore to New Orleans Coastwise Slave Trade, 1820–1860." *Slavery & Abolition* 41, no. 2 (April 2, 2020): 275–303. https://doi.org/10.1080/0144039X.2019.1660509.

Acknowledgements

This work is supported in part by the National Science Foundation under Grant No. SES-2338428 and DMS-2330089. The ideas in this work are representative of the authors and not of the NSF.