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

Hannah Butler & Andee Kaplan

Colorado State University

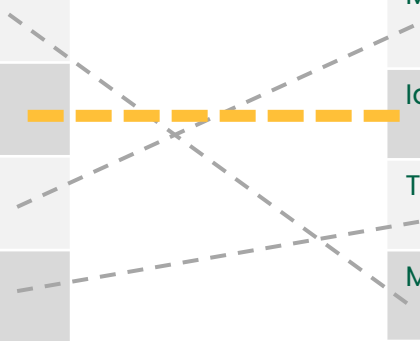
Overview

- What is Record Linkage?
- Motivation & Goal
- Record Linkage and Aliases
- Introduce Model
- Application
- Future Work

What is Record Linkage?

A method used to identify records from different datasets belonging to a common entity.

Author	Title	Publ. Date
Stephen Jay Gould	The Mismeasure of Man	1996
Ian Hacking	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	The Taming of Chance	1990
Theodore M. Porter	The Rise of Statistical Thinking: 1820-1900	2020
Michael E. Staub	The Mismeasure of Minds	2018

- **Deterministic record linkage**
  - Define rules to match field values
  - Rules can be extensive and complex
- **Probabilistic record linkage**
  - Statistical model to estimate matches between records
  - More flexibility for error-prone data
  - Allows for uncertainty quantification

- **Comparing values**
  - Values compared with a chosen similarity metric
  - Pairwise comparison is discretely categorized
  - Multinomial-mixture likelihood
- **Assumption**
  - There are no duplicates within a single file

Data & Motivation

Data on Black Americans in the 1800s

1 Fugitive Slave Advertisements

- 32,000+ ad entries
- Partially crowdsourced
- Contained in an online database run by Freedom on the Move Project ©

2 Domestic Trafficking Ship Manifests

- 63,000+ records
- aggregated and digitized by the Kinfolkology™ team lead by Jennie K. Williams, Ph.D.

3 Notes of New Orleans Slave Sales

- 1,417 records
- aggregated and digitized by the Kinfolkology™ team

4 Narrative Accounts

- 995 fugitive accounts
- written and collected by William Still
- Dataset transcribed and compiled by Historian Nick Sacco

1

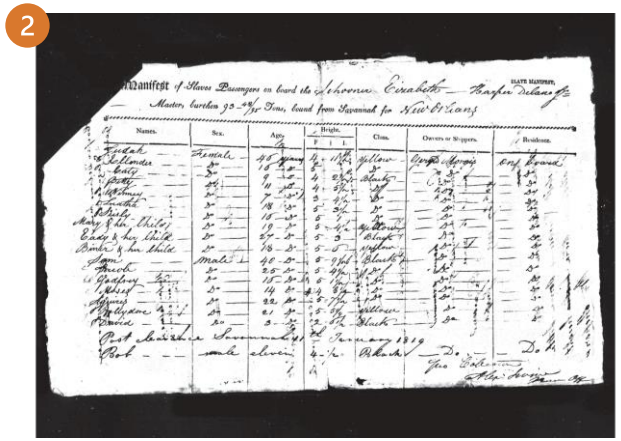
Wilmington, Dec. 16th, 1836. 1-f.

$50 Reward.

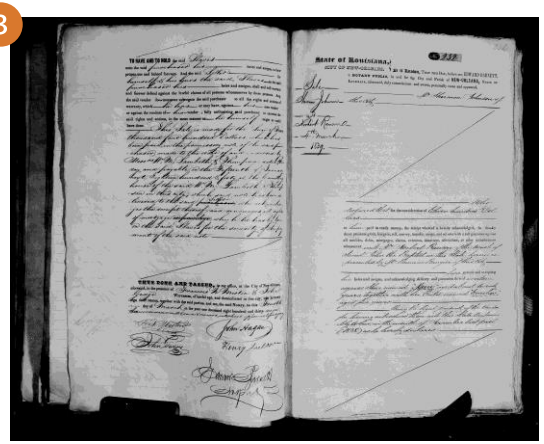
RANAWAY from the Subscriber, near Chapel Hill, on January 27th last, a Negro Fellow, named Ames, about 22 years of age. He is about 5 feet, 9 inches in height, dark complexion, slender built, stoops a little forward in walking, speaks low, but tolerably quick when spoken to; his upper teeth on the right side are somewhat defective. When last heard of he was on the road to Wilmington, where I purchased him of John W. Hamilton, on the 1st day of November last. He took with him a small Bay Mare, with blazed face, one white foot, and hip-shot in her left hip.—The above reward will be given to any person who deliver him to me or secure him in such way that I may get him again.

BURRIUS CHEEK,
Chapel Hill, Orange Co., N. C.
February 10th, 1837. 5 t-f.

2



3



4

MARY EPPS, ALIAS EMMA BROWN-JOSEPH AND ROBERT ROBINSON

A SLAVE MOTHER LOSES HER SPEECH AT THE SALE OF HER CHILD—BOB ESCAPES FROM HIS MASTER, A TRADER, WITH $1500 IN NORTH CAROLINA MONEY

Mary fled from Petersburg and the Robinsons from Richmond. A fugitive slave law-breaking captain by the name of B., who owned a schooner, and would bring any kind of freight that would pay the most, was the conductor in this instance. Quite a number of passengers at different times availed themselves of his accommodations and thus succeeded in reaching Canada.

His risk was very great. On this account he claimed, as did certain others, that it was no more than fair to charge for his services—indeed he did not profess to bring persons for nothing, except in rare instances. In this matter the Committee did not feel disposed to interfere directly in any way, further than to suggest that whatever understanding was agreed upon by the parties themselves should be faithfully adhered to.

Why Record Linkage?

- Highly fractured historical data on African Americans
- Little-no intermingling of existing databases
- Less access to knowledge of history and kin for descendants

Challenges:

- Large amounts of missing data
- Inconsistent information within and across sources
- Heterogeneous data

Silver Lining:

- **There are many known duplicates with additional information**

Aliases

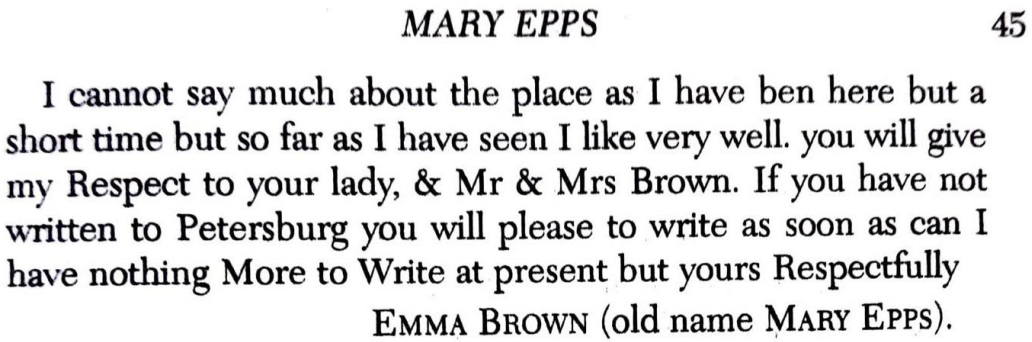
Alias: The occurrence of one or more duplications of an entity within a file, not due to error, but rather due to a known alternative piece of information

Entity	Handle	Name	Account	Business
User 1	@colorwired	Abby Joy	Personal	NA
	@mysky.ceramics	My Sky Ceramics	Business	Artist
User 2	@photogeanic	Emma Jean	Personal	NA
	@leafy_jean	E. Jean	Business	Grocery Store
User 3	@oscarbutlerphoto	Oscar	Personal	NA
	@oscarbutlermusic	Oscar Butler	Personal	NA
	@gettingfitwithoscar	Oscar Butler	Personal	NA

- Aliases give important info about user
- Info in one may be misleading/incorrect
- Repeated values can reinforce record info

Sources of Aliases in US Slave Data

- Multiple enslavers
- Relocation
- Name inconsistencies
  - Not recording full names of enslaved individuals
  - Forced name changes
  - Voluntary name changes



MARY EPPS 45

I cannot say much about the place as I have ben here but a short time but so far as I have seen I like very well. you will give my Respect to your lady, & Mr & Mrs Brown. If you have not written to Petersburg you will please to write as soon as can I have nothing More to Write at present but yours Respectfully

EMMA BROWN (old name MARY EPPS).

From *The Underground Railroad*

Goal:

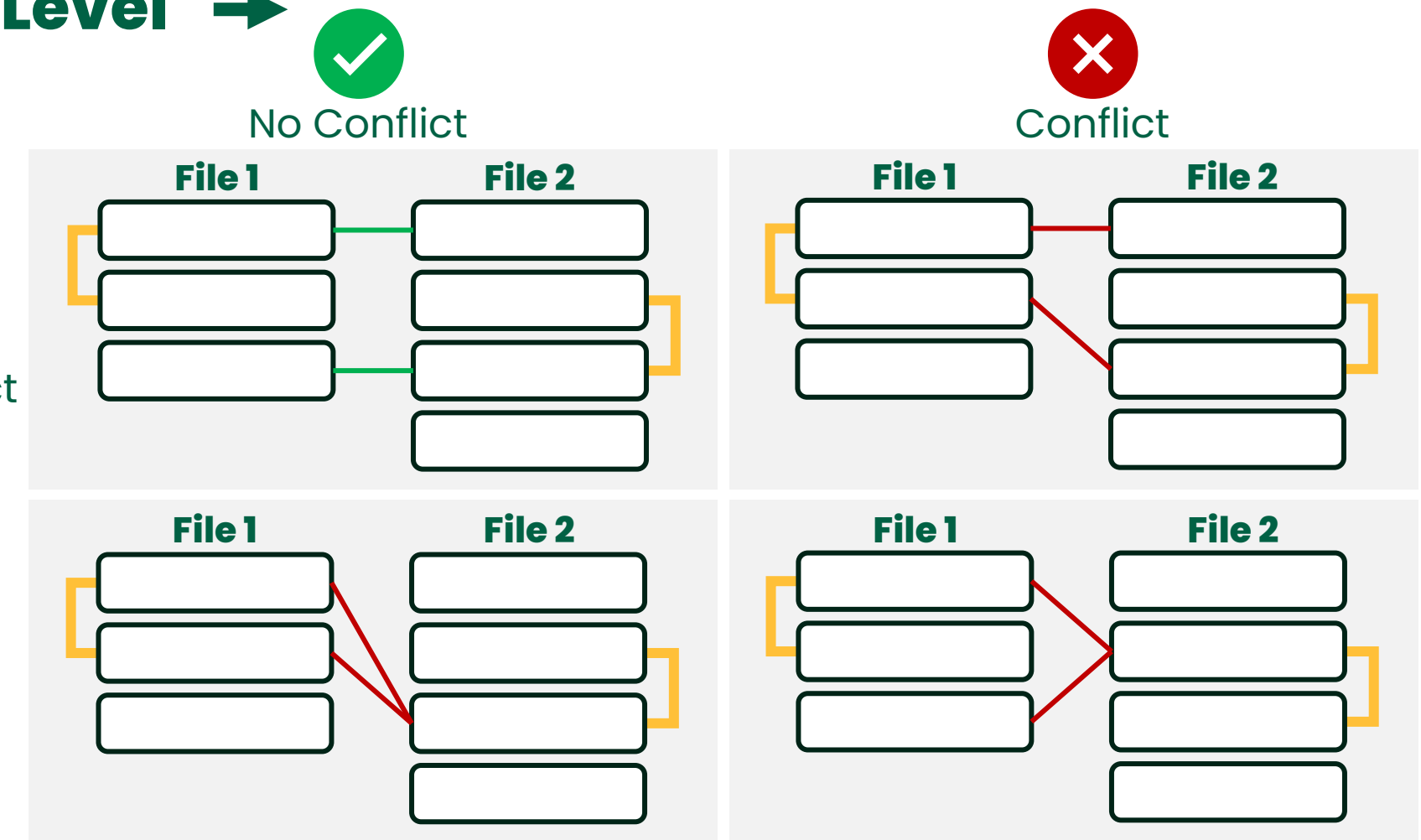
Develop an effective probabilistic record linkage model for African American genealogy that leverages the presence of alias information.

Record Linkage With Aliases

Link Conflicts with Aliases

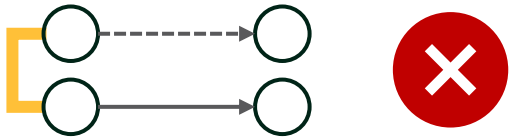
Entity Level →

Record Level ↓

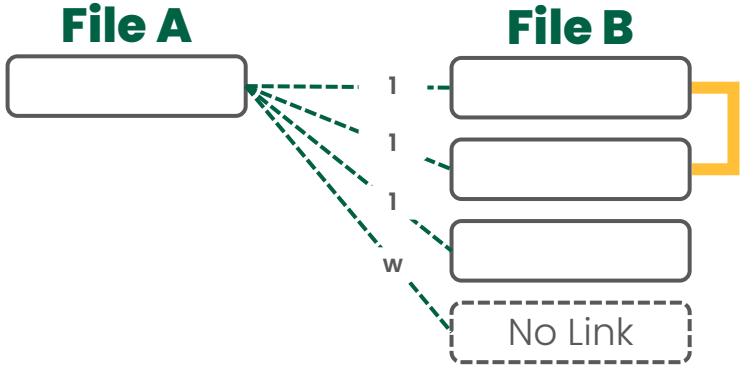


BRLwA Methodology

1. **Specify a Bayesian BRL model that accommodates alias records (BRLwA) without introducing conflict**



2. **Choose a linkage prior for Z that weights links between any two records equally**



3. **Use MCMC to generate posterior samples**

```
1.  Initialize agreement probabilities and links

2.  For each iteration, t = 1, . . ., T:
      1.  Sample Z given data, m, u
      2.  Sample m, u given data, Z
```

4. **Estimate links directly from MCMC posterior samples or consolidate records to obtain point estimate**

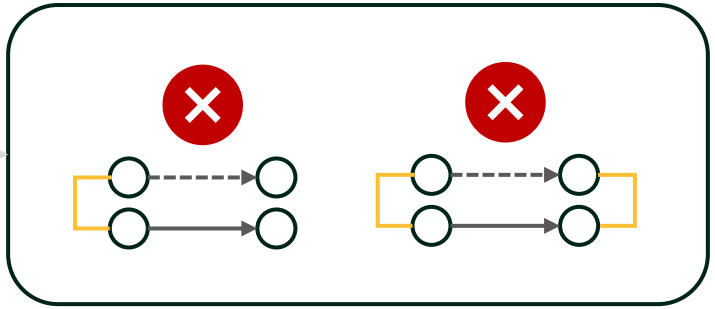
$$\widehat{\boldsymbol{Z}}_i = \begin{cases} j & \text{if } P(\boldsymbol{Z}_i = j \mid \boldsymbol{\Gamma}) > 1/2 \\ N_2 + i & \text{otherwise} \end{cases}$$

BRLwA Model

Likelihood

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

$N_1 < N_2$



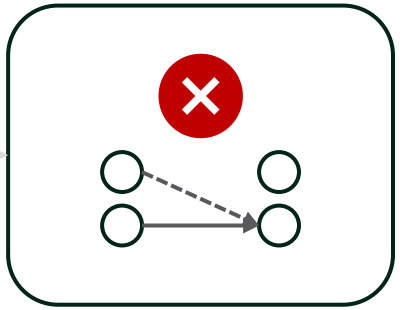
Entity-level link conflicts are prevented by V in the likelihood

Agreement level probabilities

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

Linkage vector

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

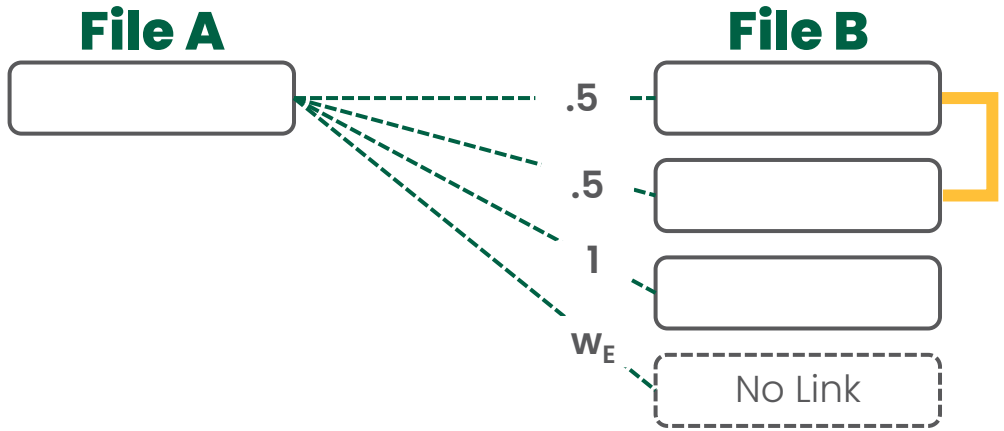


Record-level link conflicts are prevented by the prior

Choosing a Linkage Prior

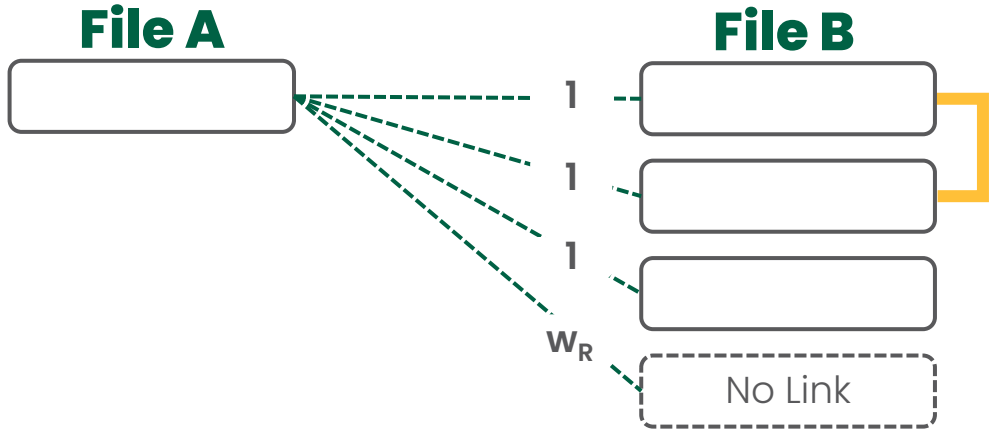
Weight by Entity (EWP)

- Each **entity** has an equal chance of being linked a priori
- **Prior probability of any conflicting link** is set to 0



Weight by Record (RWP)

- Each **record** has an equal chance of being linked a priori
- Entities with multiple aliases have a greater probability of being linked a priori
- Prior probability of record-conflicting links is set to 0
- **Likelihood of entity-conflicting links** is set to 0

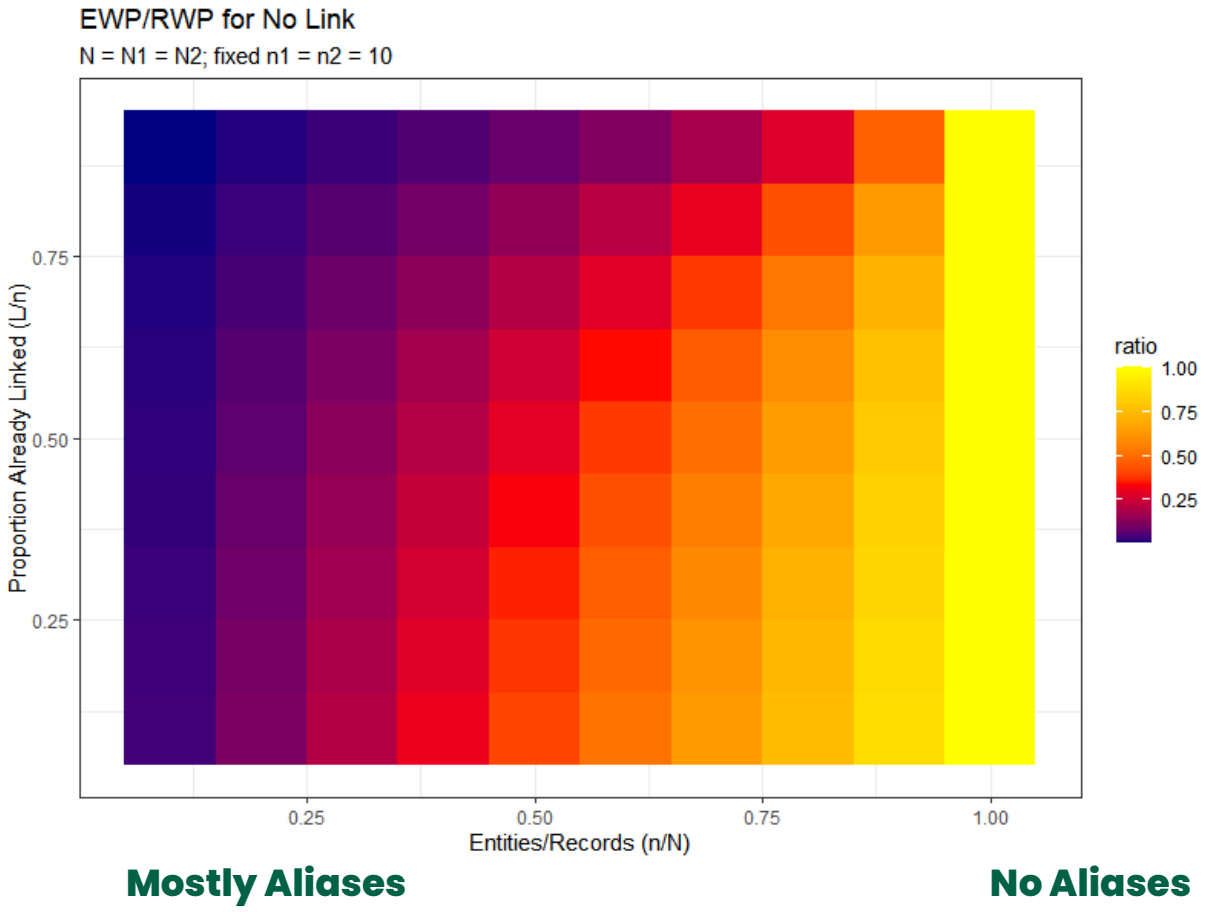


$$W_E \leq W_R$$

Prior Ratio of No Link

- EWP/RWP
  - = 1 when no alias records are present
- Simulation
  - Posterior samples show lower precision with EWP
  - More false positives with EWP

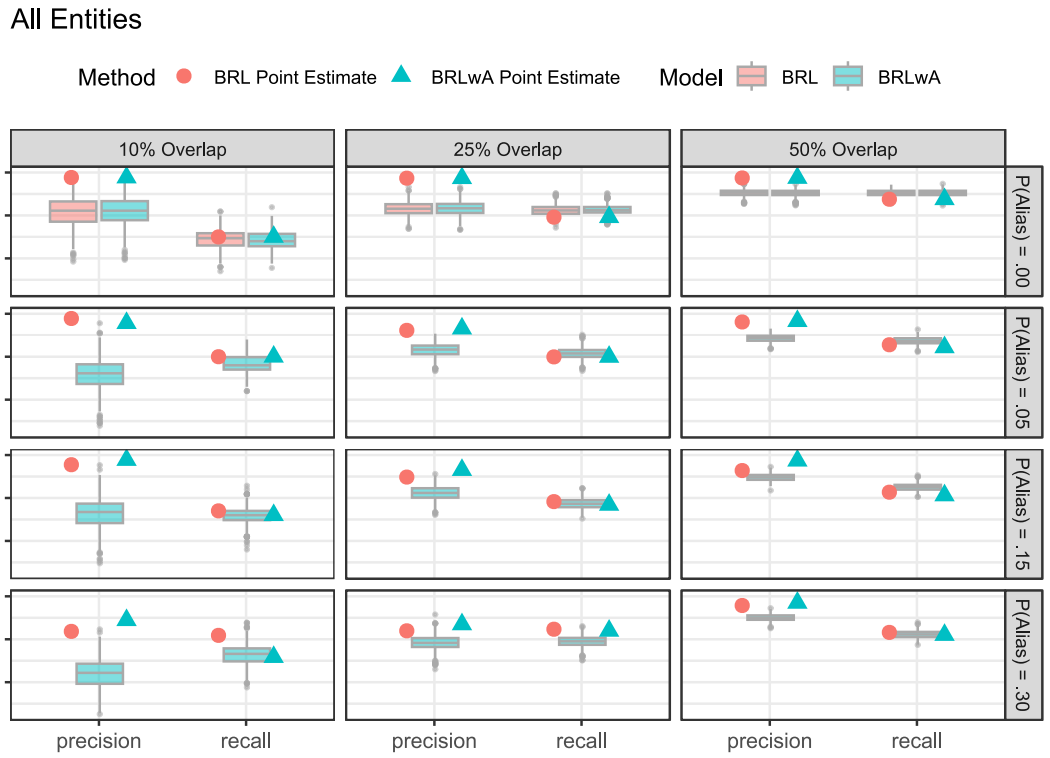
Prior probability ratio for no link for a record EWP/RWP

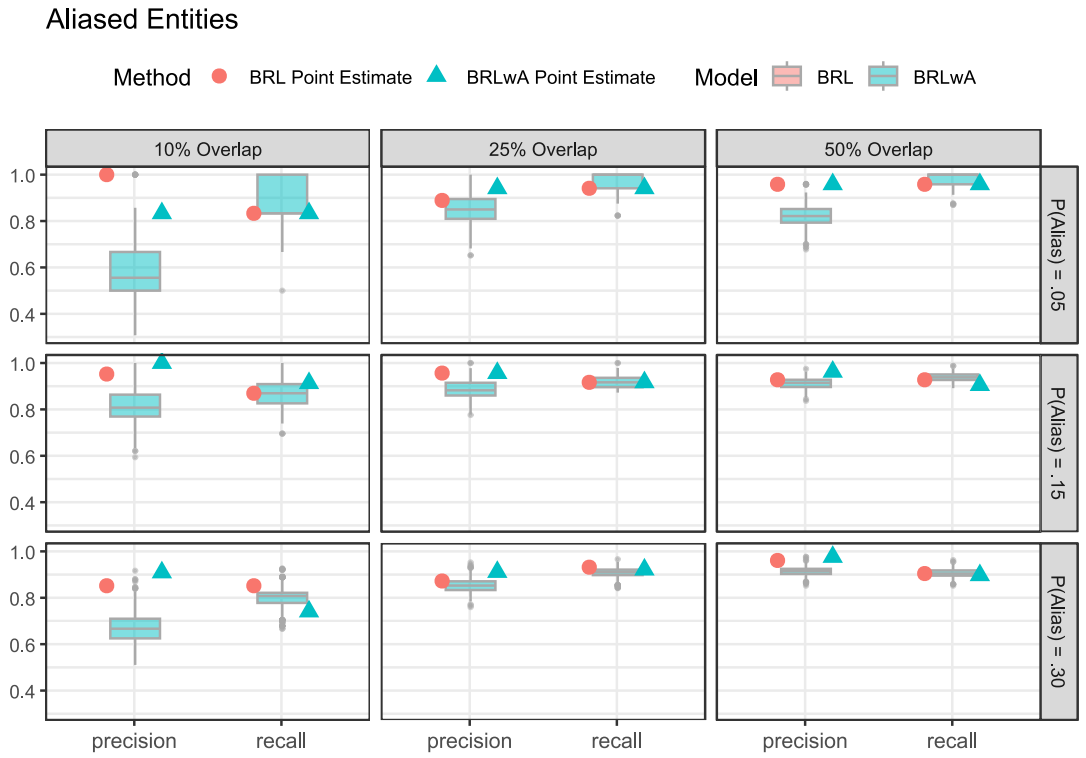


Mostly Aliases **No Aliases**

Model Performance Simulation

- Compared with post-processed estimate from original Sadinle model (BRL).
- No distribution for BRL when P(Alias) > 0





Finding Links

- **Sampling**
  - MCMC samples directly from full conditionals
- **Estimating**
  - Bayes estimate with squared or absolute error loss function
  - Posterior samples require no post-processing for conflict resolution

$$\widehat{\boldsymbol{Z}}_i = \begin{cases} j & \text{if } P(\boldsymbol{Z}_i = j \mid \boldsymbol{\Gamma}) > 1/2 \\ N_2 + i & \text{otherwise} \end{cases}$$

  - Consolidate posterior linkage vector by entity

Element of $\boldsymbol{Z}$	MCMC Iteration (minus burn-in) 1	2	3	4	5	6
1	j	j	N_2+1	N_2+1	N_2+1	N_2+1
2	N_2+2	N_2+2	N_2+2	j	j	j

$$\widehat{\boldsymbol{Z}}_1 = N_2 + 1$$
$$\widehat{\boldsymbol{Z}}_2 = N_2 + 2$$

$$\widehat{\boldsymbol{Z}}_{[1]} = [j]$$

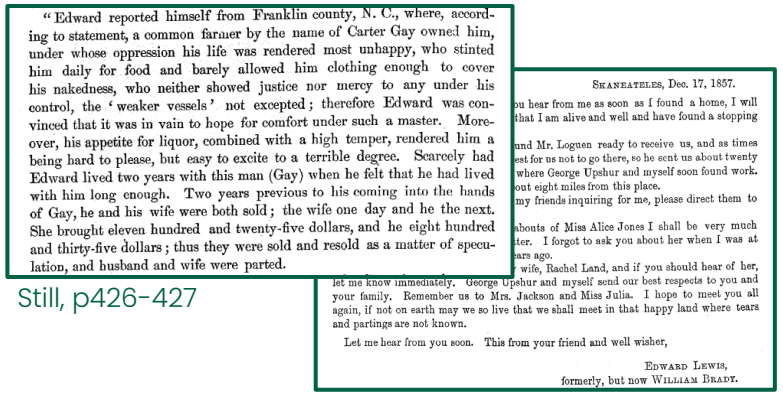
Application

Linking Runaway Slaves

- **William Still accounts X Freedom on the Move**
  - P(Alias) ~ 0.23; 1,076 of 4,584 entities.
  - estimated ~ 0.1% overlap; 5 links by BRLwA.
- **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.



"Edward reported himself from Franklin county, N. C., where, according to statement, a common farmer by the name of Carter Gay owned him, under whose oppression his life was rendered most unhappy, who stinted him daily for food and barely allowed him clothing enough to cover his nakedness, who neither showed justice nor mercy to any under his control, the 'weaker vessels' not excepted; therefore Edward was convinced that it was in vain to hope for comfort under such a master. Moreover, his appetite for liquor, combined with a high temper, rendered him a being hard to please, but easy to excite to a terrible degree. Scarcely had Edward lived two years with this man (Gay) when he felt that he had lived with him long enough. Two years previous to his coming into the hands of Gay, he and his wife were both sold; the wife one day and he the next. She brought eleven hundred and twenty-five dollars, and he eight hundred and thirty-five dollars; thus they were sold and resold as a matter of speculation, and husband and wife were parted.

SKANEATELES, Dec. 17, 1857.

...u hear from me as soon as I found a home, I will ...that I am alive and well and have found a stopping ...

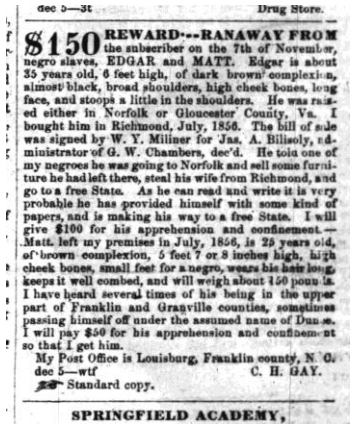
...und Mr. Loguen ready to receive us, and as times ...est for us not to go there, so he sent us about twenty ... where George Upshur and myself soon found work. ...out eight miles from this place. ...my friends inquiring for me, please direct them to ...

...abouts of Miss Alice Jones I shall be very much ...ter. I forgot to ask you about her when I was at ...ars ago. ... wife, Rachel Land, and if you should hear of her, let me know immediately. George Upshur and myself send our best respects to you and your family. Remember us to Mrs. Jackson and Miss Julia. I hope to meet you all again, if not on earth may we so live that we shall meet in that happy land where tears and partings are not known.

Let me hear from you soon. This from your friend and well wisher,

EDWARD LEWIS,
formerly, but now WILLIAM BRADY.

Still, p426-427



dec 5—3t Drug Store.

$150 REWARD---RANAWAY FROM the subscriber on the 7th of November, negro slaves, EDGAR and MATT. Edgar is about 35 years old, 6 feet high, of dark brown complexion, almost black, broad shoulders, high cheek bones, long face, and stoops a little in the shoulders. He was raised either in Norfolk or Gloucester County, Va. I bought him in Richmond, July, 1856. The bill of sale was signed by W. Y. Milliner for Jas. A. Bilisoly, administrator of G. W. Chambers, dec'd. He told one of my negroes he was going to Norfolk and sell some furniture he had left there, steal his wife from Richmond, and go to a free State. As he can read and write it is very probable he has provided himself with some kind of papers, and is making his way to a free State. I will give $100 for his apprehension and confinement.—Matt. left my premises in July, 1856, is 25 years old, of brown complexion, 5 feet 7 or 8 inches high, high cheek bones, small feet for a negro, wears his hair long, keeps it well combed, and will weigh about 150 pounds. I have heard several times of his being in the upper part of Franklin and Granville counties, sometimes passing himself off under the assumed name of Dunn. I will pay $50 for his apprehension and confinement so that I get him.

My Post Office is Louisburg, Franklin county, N. C.
dec 5—wtf C. H. GAY.
Standard copy.

SPRINGFIELD ACADEMY,

Freedom on the Move

Future Work

Future Work

- **Data Processing**
  - Further processing on string-valued variables
  - Derive additional variables from raw data
- **Model Validation**
  - Test/Validate model on data from Kinfolkology
  - Identify model shortcomings
- **Similarity Metrics**
  - Consider new ways to compare historical location data.
- **Computational intensity**
  - Incorporate hashing techniques
- **Estimation Methods**
  - Decision-theoretic estimation

Acknowledgements

- National Science Foundation[1]
- Jennie K. Williams & Kinfolkology

[1] This material is based on work 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

Sources

- 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/.
- Google Docs. “William Still Underground Railroad Data.Xlsx.” Accessed November 5, 2024. https://docs.google.com/spreadsheets/d/e/2PACX-1vTpXEcS1rao3f-ZHqeOApPoBXju6-Xeg902ujGjuDqkE_cxd0MGl8P-DFv4SMPqbg/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/.
- 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.