A Bayesian Approach to Linking Historical Records of America's Enslaved

Hannah Butler & Andee Kaplan

Colorado State University

Overview

Goal: Develop an effective probabilistic record linkage model that leverages the presence of alias information in historical slave records.

- Define Record Linkage
- Record Linkage for Historical Slave Records
- Model Specification & Method
- Simulation Performance
- Application & Results

What is Record Linkage?

A method used to identify co-referent entities from overlapping datasets.

Author	Title	Publ. Date
Stephen Jay Gould	**The Mismeasure of Man**	1996
Ian Hacking	**Taming of Chance**	**1990**
Stephen T. Ziliak, Deirdre N. McCloskey	The Cult of Statistical Significance	2008
Theodore Porter	Trust in Numbers	**2020**

Author	Title	Publ. Date
Ruha Benjamin	Race After Technology	2019
Ian Hackng	**The Taming of Chance**	**1990**
Theodore M. Porter	The Rise of Statistical Thinking	**2020**
Michael E. Staub	**The Mismeasure of Minds**	2018

- **Deterministic Record Linkage**
  - Define rules to match field values
  - Eg: `JOIN` in SQL
- **Probabilistic Record Linkage**
  - Statistical model used to estimate matches between records
  - More flexibility for error-prone data
  - **Eg: Bayesian Bipartite Record Linkage (BRL)** *(Sadinle 2017)*

Comparison Data

- Calculate distances between values of a field with a chosen similarity metric
- Classify distances as **levels of disagreement** *(Fellegi & Sunter, 1969)*
- Level of disagreement | field ~ Multinomial Mixture

Assumption

- **There are no duplicate records within a single file** *(Sadinle, 2017)*

Challenges of Linking Slave Records

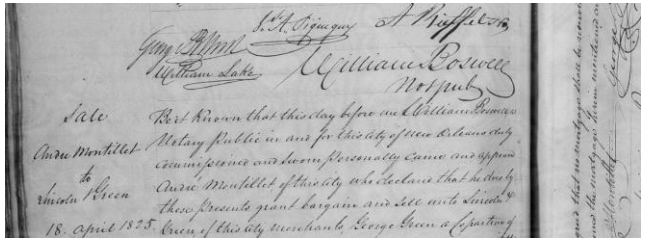
- **Data Availability**
  - Comparatively few digitized data sources
  - High rates of missing data
  - Eg: Last names
- **Uncertainty**
  - Data collection is unregulated/non-standardized
  - Falsified/contentious/subjective values
  - Transcription errors
  - Eg: Ages
- **Duplicate Records (Non-Erroneous)**
  - **Many entities have multiple variable values for a field (Aliases)**
  - Potential violations of bipartite linkage structure

From: ***Louisiana Kindred (Kinfolkology)***

	ID (LK)	Enslaved Name	Racial Descriptor	Age	Kin Relation	Kin Name
1 entity, 1 alias	**PELK100940**	Cook, James	"mulatto"	25	<NA>	<NA>
	PELK100498	Lucy	"negro"	37	Mother of	**Harrison**
1 entity, 3 aliases	**PELK100498**	Lucy	"negro"	37	Mother of	**Arthennise**
	PELK100498	Lucy	"negro"	37	Mother of	**[unnamed child]**

From: ***Oceans of Kinfolk (Kinfolkology)***

ID (OOK)	Enslaved Name	Racial Descriptor	Age	Kin Relation	Kin Name
OOK104652	Cook, James	"yellow"	20	<NA>	<NA>
OOK106770	Lewis, Lucy	"black"	35	<NA>	<NA>
OOK109450	Lucy	"dark"	40	<NA>	<NA>
OOK115546	Lucy	"black"	38	Mother of	<NA>



- Louisiana Kindred, Kinfolkology

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).

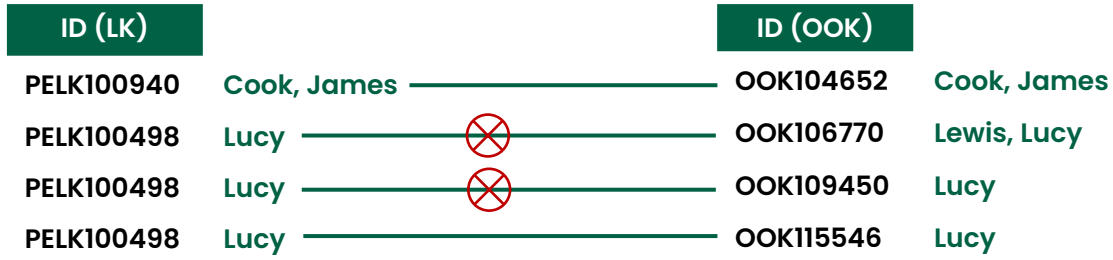
- The Underground Railroad

Existing Approaches

- **Record Linkage with Naïve Pre-Processing**
  - Remove alias values/records during data processing
  - Perform record linkage as usual

ID (LK)	Enslaved Name	Racial Descriptor	Age	Kin Relation	Kin Name
PELK100940	Cook, James	"mulatto"	25	*<NA>*	*<NA>*
PELK100498	Lucy	"negro"	37	Mother of	**Harrison**
~~**PELK100498**~~	~~Lucy~~	~~"negro"~~	~~37~~	~~Mother of~~	~~**Arthennise**~~
~~**PELK100498**~~	~~Lucy~~	~~"negro"~~	~~37~~	~~Mother of~~	~~**[unnamed child]**~~

- **Record Linkage with Post-Processing**
  - Keep all records and values
  - Perform record linkage as usual
  - Resolve conflicts after estimation



Bayesian

Bipartite Alias Record Linkage (BARL)

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

Comparisons

$$\gamma_{ij}^{f} \mid \boldsymbol{m}_f, \boldsymbol{u}_f, \boldsymbol{Z} \overset{\text{iid}}{\sim} \text{Mult}(1, K_f, \boldsymbol{m}_f^{\boldsymbol{Z}_i=j} \boldsymbol{u}_f^{\boldsymbol{Z}_i \neq j})$$

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

Level of Disagreement Probabilities

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

2. Choose a linkage prior for $\boldsymbol{Z}$ that equally weights links between any two **records**

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

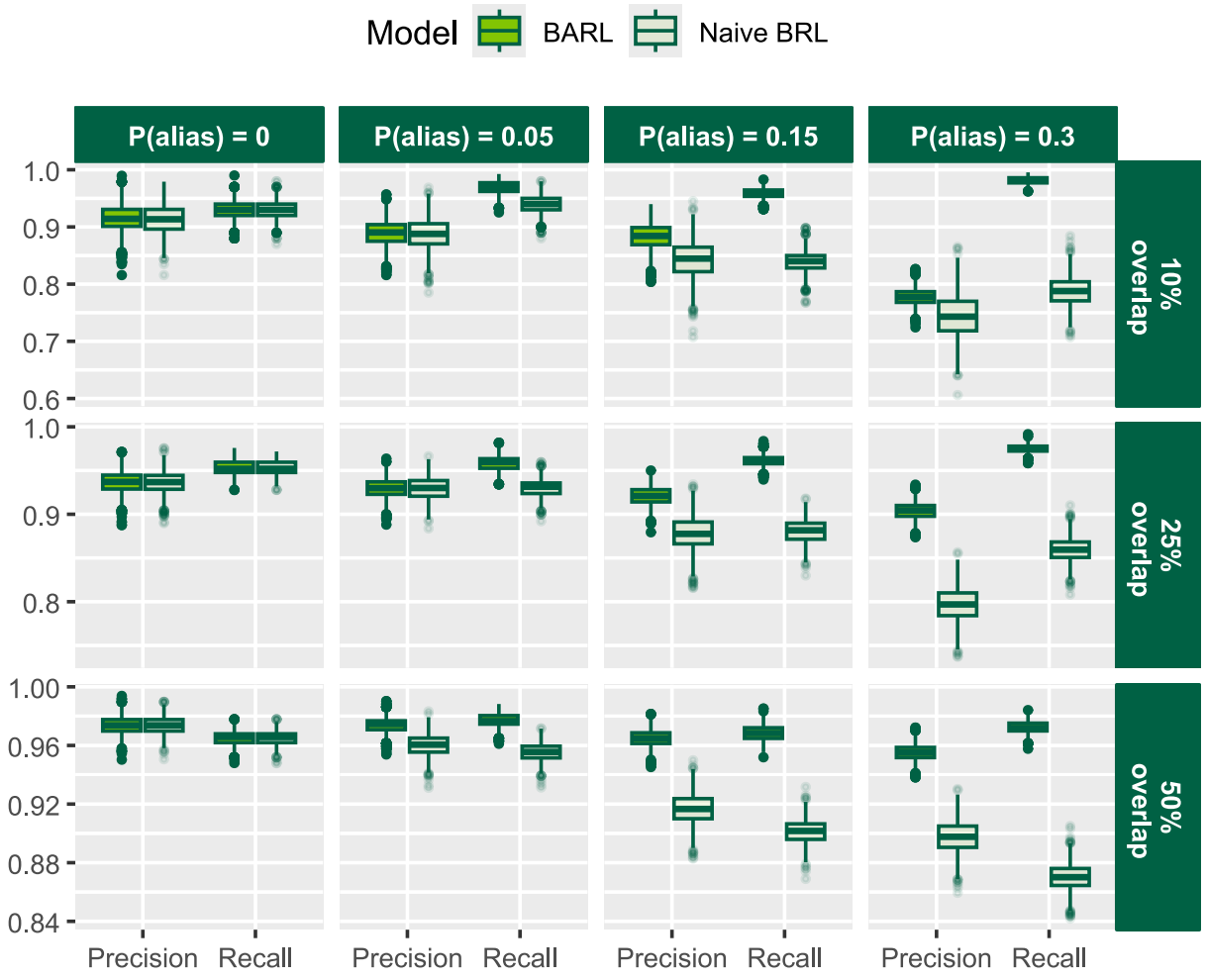
3. Use MCMC to generate posterior samples

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

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

Model Performance Simulation



Precision = TP/(TP + FP) Recall = TP/(TP + FN)

Data Simulation

- Records generated using GeCo
- Simulated overlapping datasets with aliases
- BARL compared to BRL with additional aliases removed (**Naïve BRL**)
- When no aliases are present, BARL = BRL

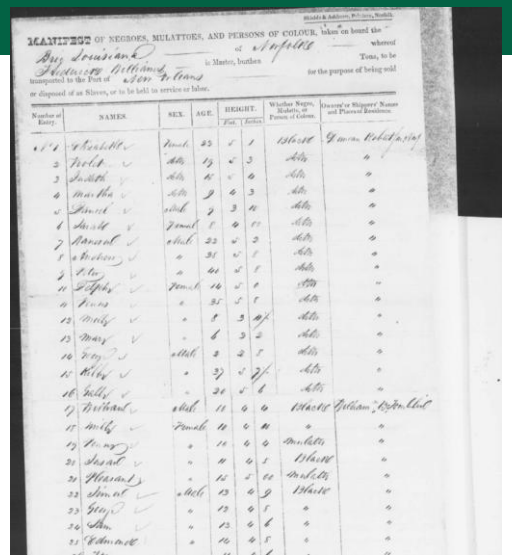
Results

- BARL: Greater average precision & recall than Naïve BRL

Applying BARL to *Kinfolkology* Data

Oceans of Kinfolk

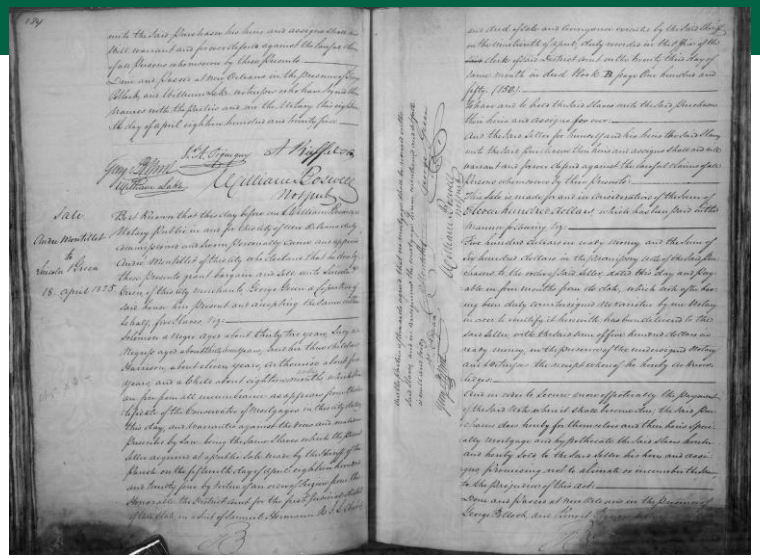
> 20,000 records of enslaved individuals trafficked by ship between domestic ports in the U.S. – transcribed & digitized


MANIFEST OF NEGROES, MULATTOES, AND PERSONS OF COLOUR, taken on board the

From: *Oceans of Kinfolk* on *Kinfolkology*

Louisiana Kindred

~ 1,500 records of individuals sold and bought in antebellum New Orleans – transcribed & digitized



From: *Louisiana Kindred of Kinfolk* on *Kinfolkology*

www.kinfolkology.org/data-overview

Aligned Variables

- Enslaved Name (first, last)
- Enslaved Age
- Enslaved Gender
- Enslaved Kin Name
- Enslaved Kin Relation
- Event Date
- Enslaver Name (first, last)
- Enslaver Location (city, county, state)

Oceans of Kinfolk

Louisiana Kindred

Linked	Embark Port
0 / 20	Unknown
0 / 144	Tampa, FL
0 / 629	St. Marks, FL
0 / 8	St. Joseph, FL
0 / 1	St. Augustine, FL
0 / 241	Savannah, GA
3 / 2332	**Richmond, VA**
0 / 3	Port Leon, FL
0 / 8	Philadelphia, PA
1 / 908	**Petersburg, VA**
0 / 431	Pensacola, FL
0 / 5	Pearlington, MS
8 / 5290	**Norfolk, VA**
0 / 5	New York, NY
0 / 117	New Smyrna, FL
1 / 982	**Mobile, AL**
0 / 45	Key West, FL
0 / 2	Jacksonville, FL
0 / 6	Hampton, VA
0 / 110	Galveston, TX
1 / 2450	**Charleston, SC**
0 / 1	Boston, MA
0 / 3	Bay of Biloxi, MS
24 / 5022	**Baltimore, MD**
0 / 310	Apalachicola, FL
0 / 1030	Alexandria, VA

Name Age at Sale

John Lewis 16
Nancy 17
Rachel 20

Nancy 2

Hetty 13
Frank 17
George 20
William Lewis 12
Louisa 16
Jane 4
Fanny 11
Easter 13

George 20

Harry 20

Betsey 19
Margaret 19
Ann *NA*
Milly 22
Kitty 17
Sarah Ann NA
Kitty 16
Henny 14
Caroline 14
Nancy 23
Levy 25
Lewis 9
Simon 20
William K. Clark 17
Moses 17
Edward Warren 20
Rebecca Anclison 20
Eliza Nichols 7
Sarah Nichols 23
Betsey 9
Priscilla 20
Mary Ann Clark 16
Catherine Johnson 18
Mary Mitchell 18

Thank You!

Acknowledgments

- National Science Foundation[1]
- Jennie K. Williams, PhD & 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.
- Kinfolkology. “Kinfolkology.” Accessed October 30, 2024. https://www.kinfolkology.org.
- National Center for Health Statistics. Division of Analysis and Epidemiology., “The Linkage of National Center for Health Statistics Survey Data to Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data (2014-2019): Matching Methodology and Analytic Considerations.”
- 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/.