Leveraging Bayesian Bipartite Record Linkage to Connect the Experiences of Victims of Antebellum Slavery

Hannah Butler

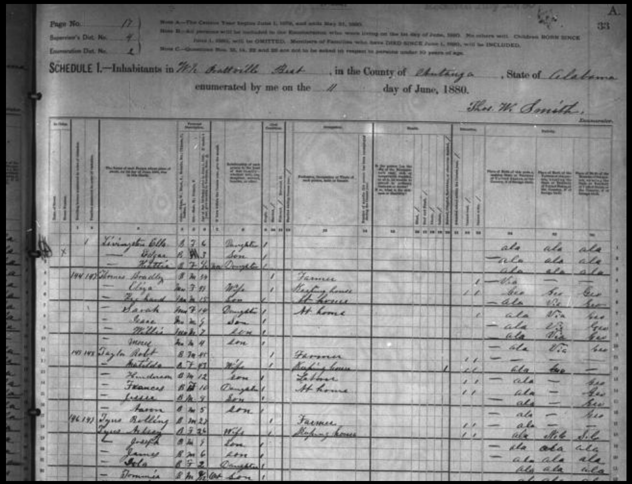
Advisor: Andee Kaplan

Committee: Kayleigh Keller, Matthew Koslovsky, Carrie Chennault (Anthropology and Geography)

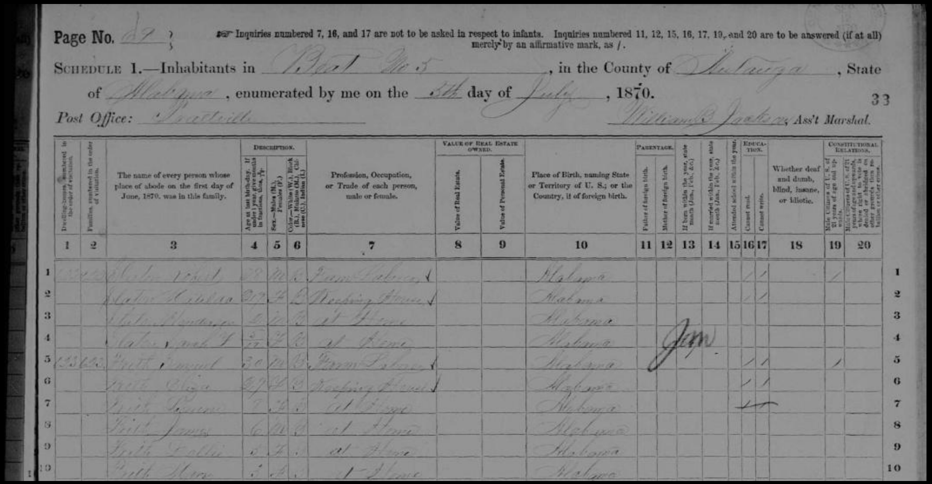
Agenda

- Introduction
- **Chapter 2** Leveraging Bayesian Bipartite Record Linkage to Connect the Experiences of Victims of Antebellum Slavery
- **Chapter 3** Visualizing Record Linkage Through the Narratives of the Enslaved
- Future Directions

1880 census, Robt Taylor



1870 census, Robert Slaton

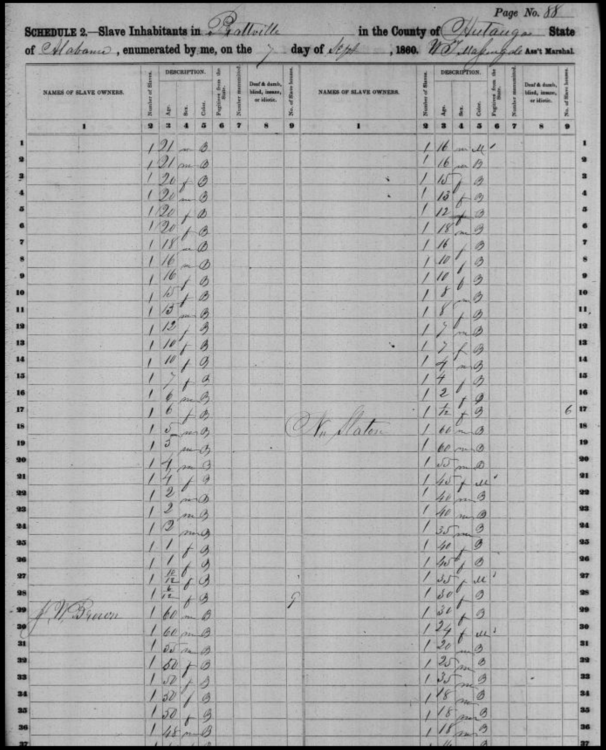

Page No.

Inquiries numbered 7, 16, and 17 are not to be asked in respect to infants. Inquiries numbered 11, 12, 15, 16, 17, 19, and 20 are to be answered (if at all) merely by an affirmative mark, as /.

SCHEDULE 1.—Inhabitants in , in the County of , State of , enumerated by me on the day of , 1870.

Post Office: Ass't Marshal.

1860 slave schedule, [*unnamed*]


SCHEDULE 2.—Slave Inhabitants in in the County of State of , enumerated by me, on the day of , 1860. Ass't Marshal.

Images from FamilySearch.org

Record Linkage

Record linkage (RL) is the process of identifying **co-referent** records from separate data sources that belong to the same entity.

1870 Census				
ID	**First Name**	**Last Name**	**Birth Year**	**County**
001	Robert	Slaton	1842	Autauga

1880 Census				
ID	**First Name**	**Last Name**	**Birth Year**	**County**
A25	Robt	Taylor	1835	Autauga
A26	Robert	Taylor	1841	Bullock

Automatic Record Linkage

Automatic RL methods exist which can systematically do* this without human oversight.

- **Deterministic RL**
  - Determines co-reference via the satisfaction of conditions/rules
  - If a pair of records satisfy all pre-specified conditions for a match, then they are linked

- **Probabilistic RL**
  - Estimates the likelihood of the data, conditional on the co-reference status between pairs of records.
  - Allows for greater flexibility and uncertainty quantification

By leveraging computational power, these methods can perform RL on digitized records much more efficiently than manual RL

- *For the remainder of this presentation, RL will refer to automatic RL unless specifically noted*

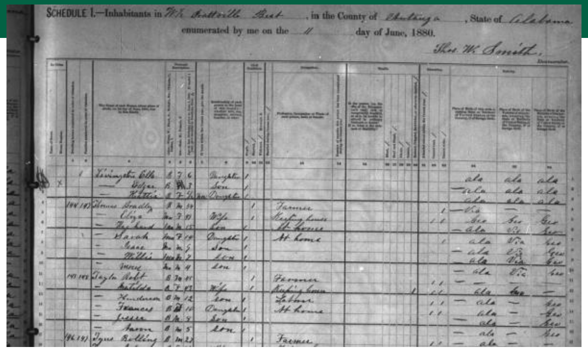
Why is Record Linkage Difficult?

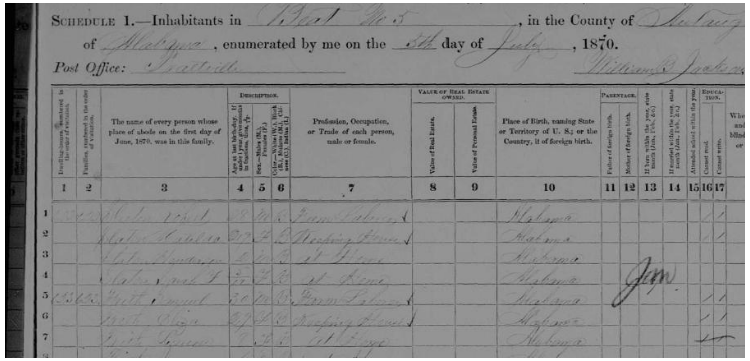
- No common unique identifier/ no ID mapping
- Inconsistency in fields (variables)
- Shared values
- Errors or estimates
- Missing values

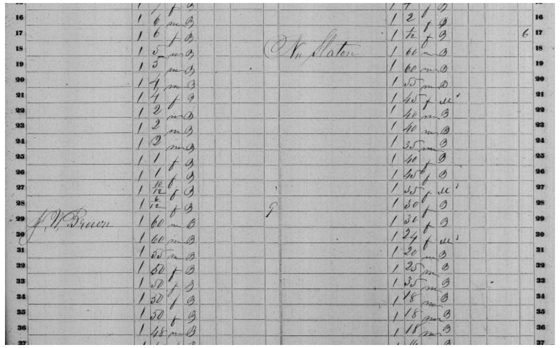
Name	Birth Year	Event Place
Robt Taylor	1835	Autauga, AL

Name	Birth Year	Event Place
Robert Slaton	1842	Autauga, AL

Name	Birth Year	Event Place
[unnamed]	1842	Autauga, AL
[unnamed]	1842	Autauga, AL
[unnamed]	1842	Autauga, AL
[unnamed]	1835	Autauga, AL







Applications of Record Linkage

Areas where record linkage has been applied are myriad and diverse

- Official Statistics (Jaro 1989; Winkler 1991, 2001; Kaplan et al. 2022)
- Public Health (Newcombe et al. 1959; Gutman et al. 2013)
- Social Networks (Sosa and Rodríguez 2023)
- Ecology (Lu et al. 2022; Drew et al. 2025)

- **Historical**

Linking census data from 1850-present (Abramitzky et al. 2021)

Linking enslavers involved in the coastwise slave trade (Steckel and Ziebarth 2013)

"Linkage based on variables that are not expected to change over time"

- **Humanitarian Efforts**

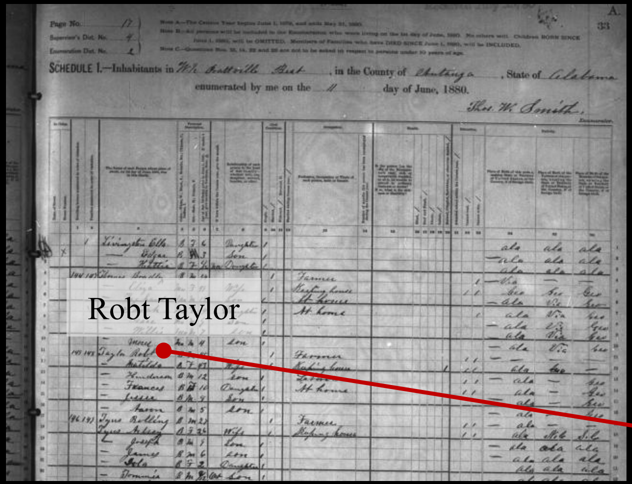
Accurately accounting for lethal violence in vulnerable populations (Sadinle 2014; Gargiulo et al. 2024)

"Especially challenging when records are subject to errors and missing values"

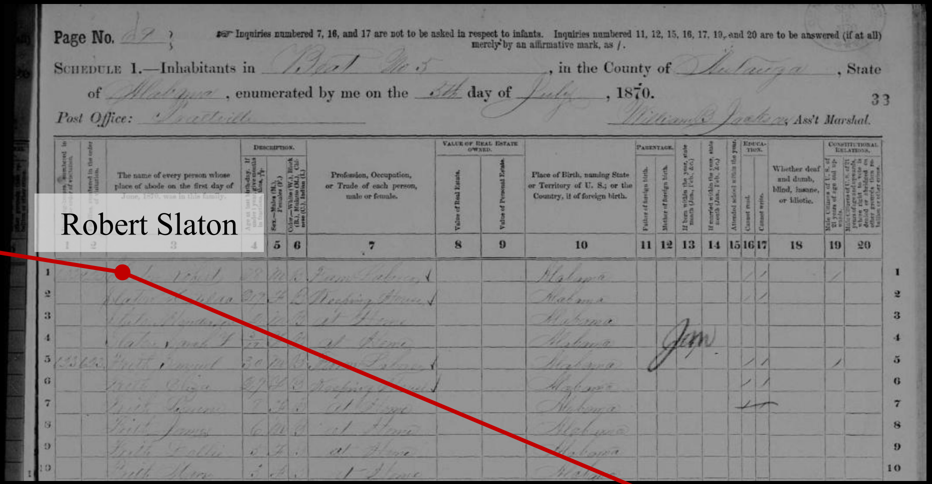
Chapter 2

Leveraging Bayesian Bipartite Record Linkage to Connect the Experiences of Victims of Antebellum Slavery

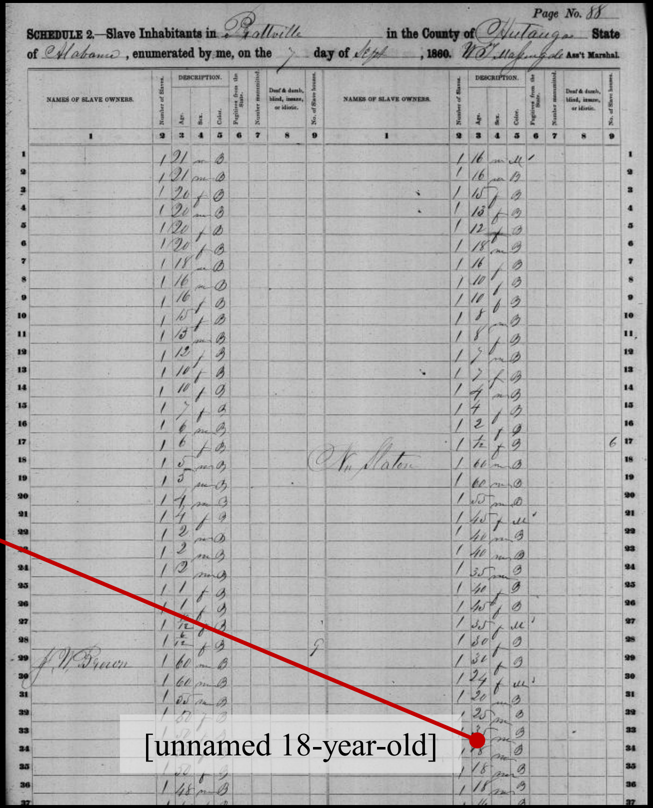
1880 census, Robt Taylor



1870 census, Robert Slaton



1860 slave schedule, [*unnamed*]



Images from FamilySearch.org

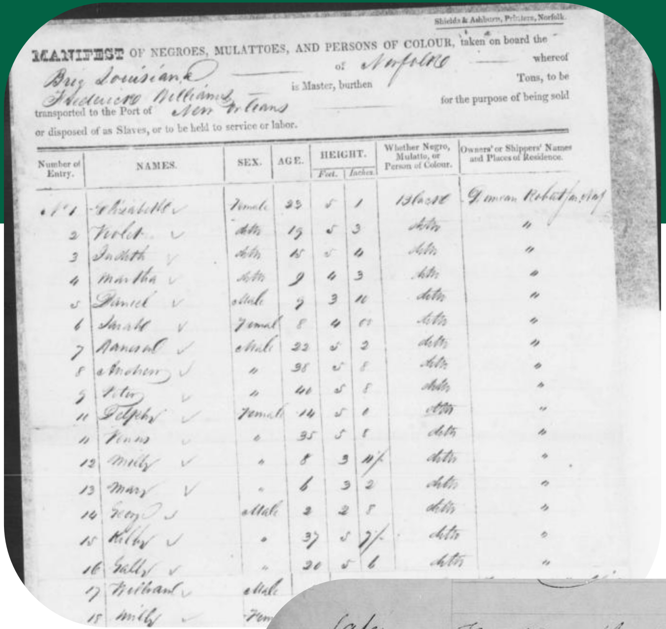
The Data

Oceans of Kinfolk

- Source: Ship manifests from domestic coastwise slave trade (1818-1860)
- ~ 21,000 enslaved persons' records

Louisiana Kindred

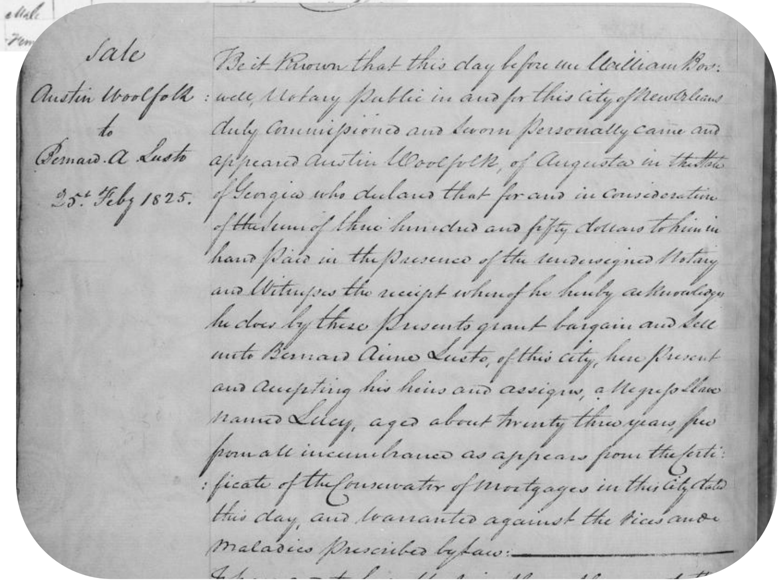
- Source: Notarial records from the sales of enslaved people in New Orleans (1811-1862)
- ~1,500 enslaved persons' records



Shields & Ashburn, Printers, Norfolk.

MANIFEST OF NEGROES, MULATTOES, AND PERSONS OF COLOUR, taken on board the Brig Louisiana of Norfolk whereof is Master, burthen Tons, to be transported to the Port of New Orleans for the purpose of being sold or disposed of as Slaves, or to be held to service or labor.

Number of Entry.	NAMES.	SEX.	AGE.	HEIGHT. Feet.	HEIGHT. Inches.	Whether Negro, Mulatto, or Person of Colour.	Owners' or Shippers' Names and Places of Residence.



Sale
Austin Woolfolk
to
Bernard A Lusto
25th Feby 1825.

Be it Known that this day before me William Bos:well Notary Public in and for this city of New Orleans duly Commissioned and Sworn personally came and appeared Austin Woolfolk, of Augusta in the State of Georgia who declared that for and in consideration of the sum of three hundred and fifty dollars to him in hand paid in the presence of the undersigned Notary and Witnesses the receipt whereof he hereby acknowledges he does by these presents grant bargain and sell unto Bernard Aimé Lusto, of this city, here present and accepting his heirs and assigns, a Negro fellow named Liley, aged about twenty three years free from all incumbrances as appears from the certificate of the Conservator of Mortgages in this city dated this day, and warranted against the vices and maladies prescribed by law:

Images from Kinfolkology.com

Why This Data?

Historian Dr. Jennie K. Williams has done the work of digitizing this data.

- Tens of thousands of people were trafficked from the upper South to New Orleans by traders where they were usually sold (Pritchett 2001; Steckel and Ziebarth 2013; Williams 2020; Jones 2021).
- Some records have already been hand-linked between *Oceans of Kinfolk* and *Louisiana Kindred*.

Data on Enslaved People

- No common unique identifier/ no ID mapping
- Inconsistency in fields (variables)
- Shared values
- Errors or estimates
- Missing values

+

- Source materials fragmented and scattered (Thomas and Fowler 2017)
- Lack of vital/official records
- Little recognition of last names
- Low literacy

Record ID (*Louisiana Kindred*)	**First Name (*Louisiana Kindred*)**	**Record Count for First Name (*Oceans of Kinfolk*)**	**Record Count for First Name & Similar Age (*Oceans of Kinfolk*)**
P-E-LK-101140	Milly	107	27
P-E-LK-100534	Margaret	107	29

How do we accurately link records with so much uncertainty?

Probabilistic Record Linkage

- **Classical Record Linkage Models**
  - Models the disagreement between the fields of record pairs (Fellegi and Sunter 1969; Sadinle 2017)
  - Requires datasets to have no internal co-reference structure (i.e. 1 record per entity in each dataset)

- **Latent Entity Models**
  - Models the fields in the data directly (Tancredi and Liseo, 2011; Steorts et al. 2016)
  - Able to link across and within datasets simultaneously
  - Requires knowledge of field distributions → Prohibitively restrictive for data on enslaved individuals

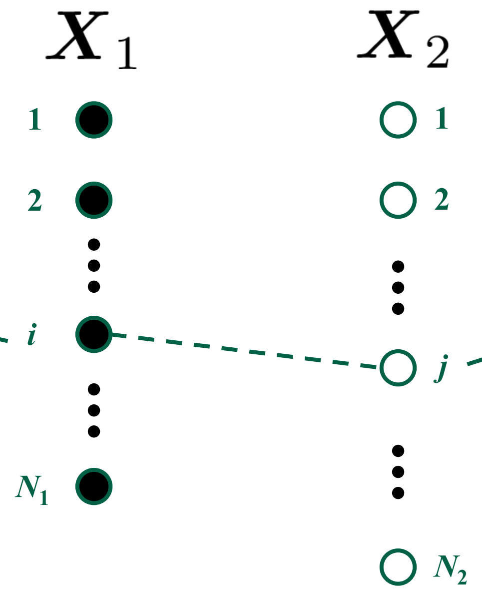
Bayesian implementations allow for better uncertainty propagation

(Fortini et al. 2001; Tancredi and Liseo 2011; Steorts et al. 2016; Sadinle 2017)

Comparison Data for Classical RL

(Fellegi and Sunter 1969): Considered estimating the links between two files with records $\boldsymbol{X}_1 = \{ i = 1, \ldots, N_1\}$ and $\boldsymbol{X}_2 = \{j = 1, \ldots, N_2\}$ by first <u>categorizing continuous measurements of discrepancy as discrete levels of disagreement for each field.</u>

ID	First Name	Last Name	Birth Year
i	Robert	Slaton	1942



ID	First Name	Last Name	Birth Year
j	Robt	Taylor	1935

- The size of disagreement, $| \,.\, |_f$, in field f is categorized into one of K_f intervals partitioning $\mathbb{R}^+ \cup \{0\}$.
- For records i in $\boldsymbol{X}_1$ and j in $\boldsymbol{X}_2$, this can be encoded in vectors $\boldsymbol{\gamma}^f_{ij}$, where $\boldsymbol{\gamma}^f_{ij}(k) = \mathbf{1}(| x_{if} - x_{jf} |_f \text{ in } I_{k-1})$.

Comparison Data for Classical RL

(Fellegi and Sunter 1969): **Γ** is the collection of all pairwise record disagreement levels between common linkage fields in the data

ID	First Name	Last Name	Birth Year
i	Robert	Slaton	1942

ID	First Name	Last Name	Birth Year
j	Robt	Taylor	1935

First Name ($f = 1$)
(Robert, Robt)

$\gamma^{1}_{ij} = (\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0})$

$K_1 = 4$

$I_0 = \{0\}$ Perfect Agreement | $I_1 = (0, d_1]$ **A little disagreement** | $I_2 = (d_1, d_2]$ Moderate disagreement | $I_3 = (d_2, \infty)$ High disagreement

Last Name ($f = 2$)
(Slaton, Taylor)

$\gamma^{2}_{ij} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1})$

$K_2 = 4$

$I_0 = \{0\}$ Perfect Agreement | $I_1 = (0, d_1]$ A little disagreement | $I_2 = (d_1, d_2]$ Moderate disagreement | $I_3 = (d_2, \infty)$ **High disagreement**

Birth Year ($f = 3$)
(1942, 1935)

$\gamma^{3}_{ij} = (\mathbf{0}, \mathbf{1}, \mathbf{0})$

$K_3 = 3$

$I_0 = \{0\}$ Perfect Agreement | $I_1 = (0, d_1]$ **Moderate Disagreement** | $I_2 = (d_1, \infty)$ High disagreement

Comparison Data for Classical RL

$\boldsymbol{\Gamma}$ is the collection of all pairwise record agreements/disagreements between common linkage fields in the data

Comparison for a field *f*

$$\gamma_{ij}^{f}(k) = \mathbf{1}(|x_{if} - x_{jf}|_f \in I_{k-1})$$

Comparisons for all fields,
for a record-pair (*i, j*)

$$\gamma_{ij} = \{\gamma_{ij}^{f} \mid f = 1, \ldots, F\}$$

Comparisons for all fields for a record-pair,
for all record-pairs

$$\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_{ij} \mid i = 1, \ldots, N_1, j = 1, \ldots, N_2\}$$

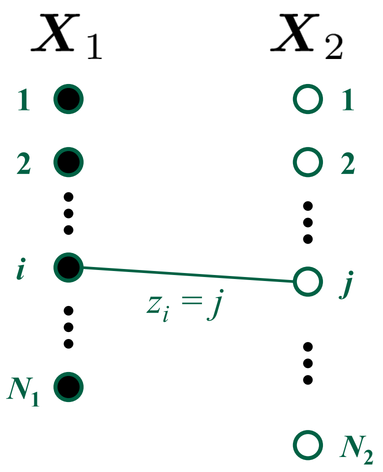
Representing Links

Links between records across datasets can be represented in different ways. Let M be the set of linked pairs:

$$M = \{(i, j) \in \boldsymbol{X}_1 \times \boldsymbol{X}_2 \mid i, j \text{ are co-referent}\}$$

Graph

Co-reference graph



Labels

Co-reference label for record i in $\boldsymbol{X}_1$

$$z_i = \begin{cases} j & \text{if } (i, j) \in M \\ N_2 + i & \text{otherwise} \end{cases}$$

Classical RL Likelihood

Model comparisons as a multinomial mixture, conditional on the link status, or label $\boldsymbol{z}_i$ (Fellegi and Sunter 1969).

- When records i and j are linked, Pr($\gamma^f_{ij}(k) = 1$) = $m_f(k)$, the kth component of $\boldsymbol{m}_f$
- When records i and j are unlinked, Pr($\gamma^f_{ij}(k) = 1$) = $u_f(k)$, the kth component of $\boldsymbol{u}_f$

$$\gamma^f_{ij} \mid \boldsymbol{z}, \boldsymbol{m}_f, \boldsymbol{u}_f \sim \text{Mult}\left(1, K_f, \boldsymbol{m}_f\right) \mathbf{1}(z_i = j) + \text{Mult}\left(1, K_f, \boldsymbol{u}_f\right) \mathbf{1}(z_i \neq j)$$

When records i and j are linked

$$\gamma^f_{ij} \mid z_i = j, \boldsymbol{m}_f \sim \text{Mult}\left(1, K_f, \boldsymbol{m}_f\right)$$

When records i and j are unlinked

$$\gamma^f_{ij} \mid z_i \neq j, \boldsymbol{u}_f \sim \text{Mult}\left(1, K_f, \boldsymbol{u}_f\right)$$

Classical RL Likelihood

Assuming that comparisons are conditionally independent and fields are independent, the full likelihood is

Ignorability indicator

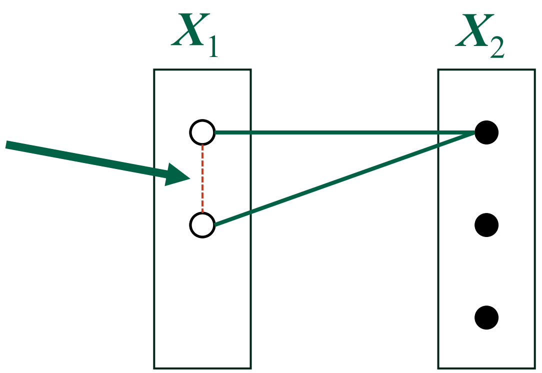
$$\mathcal{L}(\boldsymbol{\Gamma} \mid \boldsymbol{z}, \boldsymbol{m}_1, \ldots, \boldsymbol{m}_F, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_F) = \underbrace{\prod_{i=1}^{N_1} \prod_{j=1}^{N_2} \underbrace{\prod_{f=1}^{F} \underbrace{\prod_{k=1}^{K_f} \left[\boldsymbol{m}_f(k)^{\mathbf{1}(z_i=j)} \boldsymbol{u}_f(k)^{\mathbf{1}(z_i \neq j)} \right]^{\gamma_{ij}^f(k)\omega_{ij}^f}}_{\text{Probability of the observed level of disagreement in field } f \text{ between record-pair } (i,\, j) \text{, given link status}}}_{\text{across all fields}}}_{\text{for all record pairs}}$$

Ignorability: When it is assumed that a comparison is missing at random (MAR), it can be ignored, i.e., it does not add or detract from the likelihood (Little and Rubin 2002). When a comparison for record-pair (i, j) is missing in field f, $\omega^f_{ij} = 0$, otherwise $\omega^f_{ij} = 1$.

Bayesian Record Linkage

Fellegi and Sunter give us a way to test links (1969), but not an efficient way to estimate multiple links simultaneously without *transitive conflicts*.

- Transitive conflicts occur when two links imply co-reference between two distinct records.



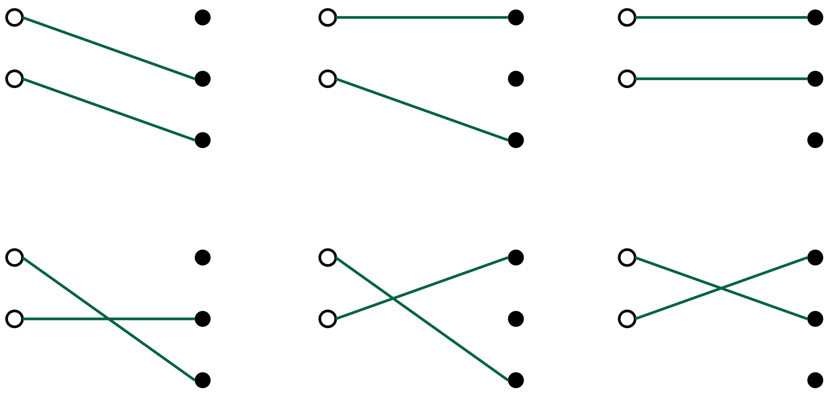
Bayesian Approach:

- A prior can be imposed on the linkage vector $\boldsymbol{z}$ to account for existing links (Fortini et al. 2001; Tancredi and Liseo 2011; Steorts et al. 2016; Sadinle 2017).

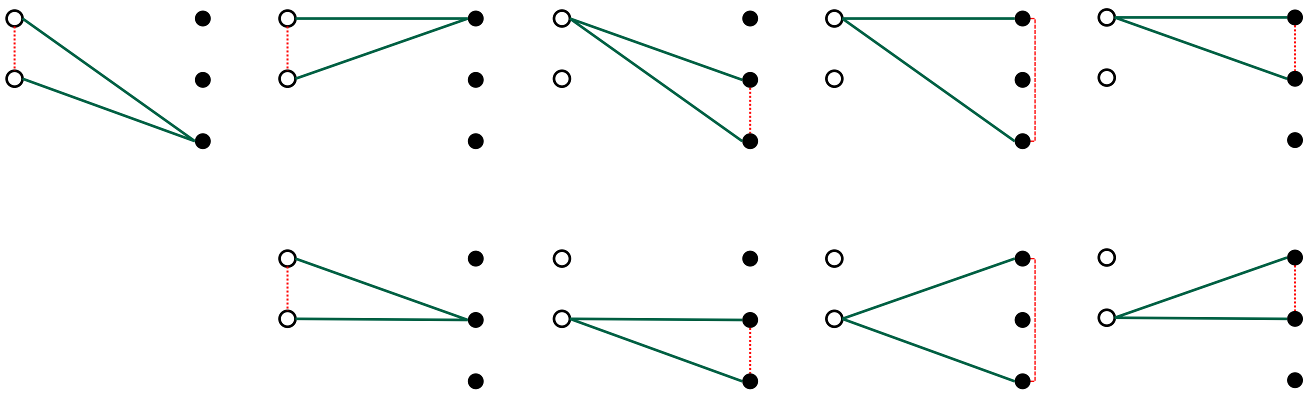
- This also allows for a direct estimate of the probability for $\boldsymbol{z} \mid \boldsymbol{\Gamma}$.

Beta-Bipartite Prior

Sadinle proposed a "Beta-Bipartite" prior for $\boldsymbol{z}$ that would prevent transitive conflicts, <u>assuming there are no co-referent records *within* a file</u> (Sadinle 2017).



P($\boldsymbol{z}$ | 2 links) = 1/6



P($\boldsymbol{z}$ | 2 links) = 0

Comparing Records of Enslaved People

$\boldsymbol{X}_1$

Louisiana Kindred

ID	Name	Relation 1	Relative 1	Relation 2	Relative 2
1	Milly	Mother of	Sarah Ann	Mother of	Hencen
2	Margaret	Mother of	*[unnamed child]*		

$\boldsymbol{X}_2$

ID	Name	Relation	Relative
A	---	---	---
B	Milly	Mother of	Sarah Ann
C	---	---	---

$\boldsymbol{X}_2$

ID	Name	Relation	Relative
A	---	---	---
B	Milly	Mother of	Hencen
C	---	---	---

Alias values for relatives

- Restrictive data structure
- Not due to error – should not be merged or removed

Comparison ambiguity

- How to compare Milly's record to other records?
- Would a comparator capable of this perform equitably in different scenarios?

Alias Records

An <u>alias record</u> or <u>alias</u> is one, of possibly multiple, distinct co-referent records with values differing non-erroneously in one or more field.

Reconstitute data into a “long” format (Wickham 2014)

- Records within a file can now be co-referent
- Fewer entities than records (3 entities, 4 records)
- Now, the numbers of entities, n_1 and n_2, are less than the numbers of records, N_1 and N_2, respectively.

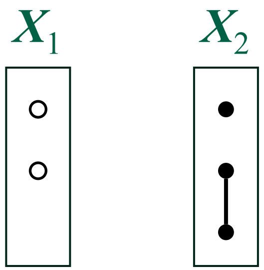
This structure violates the assumption that no co-referent records are present <u>within a single dataset</u>.

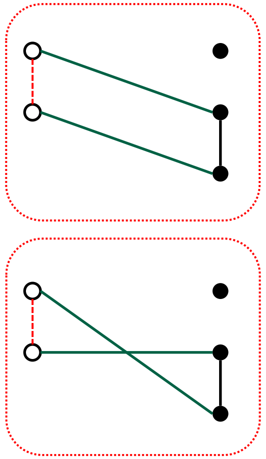
ID	Name	Relation	Relative
1	Milly	Mother of	Sarah Ann
1	Milly	Mother of	Hencen
2	Margaret	Mother of	*[unnamed child]*

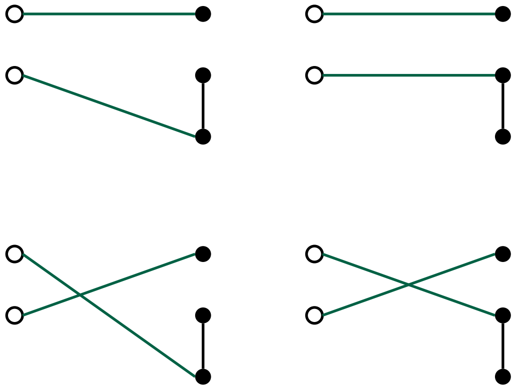
Accommodating Alias Data

- Some previously valid z now present transitive co-reference conflicts from the Beta-Bipartite prior.

➢ Multiply Beta-bipartite prior by an indicator to zero out graphs with transitive conflicts and rescale probabilities of valid graphs







P(z | 2 links)$\mathbf{1}$(z valid) = 0

P(z | 2 links)$\mathbf{1}$(z valid) = 1/4

Aliased Beta-Bipartite Prior

$$\pi \sim \text{Beta}(a, b)$$

Probability that record i in $\boldsymbol{X}_1$ is linked to a record in $\boldsymbol{X}_2$

$$\mathbf{1}(i \text{ linked}) \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)$$

Indicator of whether record i in $\boldsymbol{X}_1$ is linked

$$L = \sum_{i=1}^{N_1} \mathbf{1}(i \text{ linked})$$

Number of linked records in $\boldsymbol{X}_1$

Sadinle, 2017

Any permutation of L distinct labels from $\boldsymbol{X}_2$ is then equally likely, with probability $\dfrac{1}{\binom{N_2}{L} L!}$

Marginalizing over π, the Beta-bipartite prior for a linkage vector $\boldsymbol{z}$ is

$$P(\boldsymbol{z} \mid L, a, b) \propto \frac{B(a + L, b + N_1 - L)(N_2 - L)!}{B(a, b) N_2!} \mathbf{1}(z_i \neq z_{i'}, \forall i \neq i')$$

$\times \mathbf{1}(\boldsymbol{z} \text{ is valid})$

Aliased Beta-Bipartite Prior (A-BRL)

$$P(\boldsymbol{z} \mid L, a, b) \propto \frac{B(a+L, b+N_1-L)(N_2-L)!}{B(a,b)N_2!}\mathbf{1}(\boldsymbol{z} \text{ is valid})$$

Valid:

- $z_i \neq z_{i'}$ for all $i \neq i'$ (Sadinle, 2017)
- L is less than $\min(n_1, n_2)$
- For i linked, $z_{i'}$ is unlinked for co-referent $i' \neq i$
- For $z_i = j$, $z_{i'} \neq j'$ co-referent to j

- Maintains bipartite graphical structure between records as imposed by z from Beta-bipartite prior.
- Prevents transitive conflicts from arising due to the presence of alias records.

Aliased Bayesian Bipartite Record Linkage

The Aliased Bayesian Bipartite Record Linkage (A-BRL) model is

$$\mathcal{L}(\boldsymbol{\Gamma} \mid \boldsymbol{z}, \boldsymbol{m}_1, \ldots, \boldsymbol{m}_F, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_F) = \prod_{i=1}^{N_1} \prod_{j=1}^{N_2} \prod_{f=1}^{F} \prod_{k=1}^{K_f} \left[\boldsymbol{m}_f(k)^{\mathbf{1}(z_i=j)} \boldsymbol{u}_f(k)^{\mathbf{1}(z_i \neq j)} \right]^{\gamma_{ij}^f(k) \omega_{ij}^f}$$

$$P(\boldsymbol{z} \mid L, a, b) \propto \frac{B(a+L, b+N_1-L)(N_2-L)!}{B(a,b)N_2!} \mathbf{1}(\boldsymbol{z} \text{ is valid})$$

$$\left.\begin{aligned} \boldsymbol{u}_f \mid \boldsymbol{\beta}_f &\sim \mathrm{Dir}(\boldsymbol{\beta}_f) \\ \boldsymbol{m}_f \mid \boldsymbol{\alpha}_f &\sim \mathrm{Dir}(\boldsymbol{\alpha}_f) \end{aligned}\right\}$$

Dirichlet priors are put on the disagreement-level probabilities for conjugacy

With assumptions of conditional independence, field independence and ignorability in the likelihood.

Estimating Link Labels

The loss function specified by Sadinle (2017) is

- Additive
- Equally penalizes false negatives, false positives, false positives linked to the wrong record

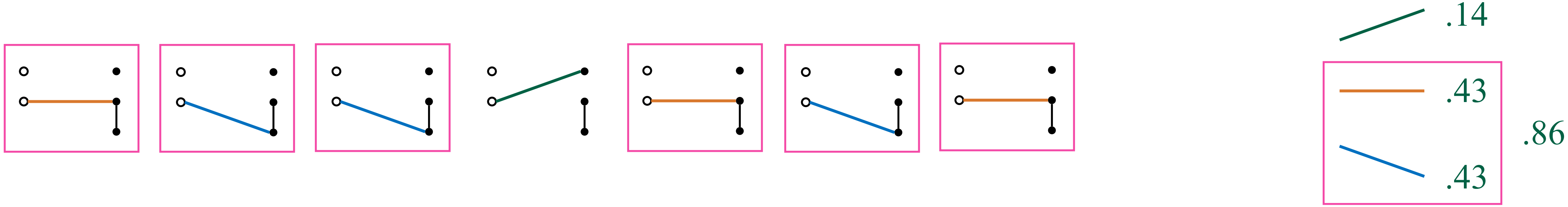
Under this loss, the Bayes estimate of the label for *i* is

$$\widehat{z}_i = \begin{cases} N_2 + i & \text{if } \sum_{j=1}^{N_2} j\mathbf{1}\left(P(z_i = j \mid \mathbf{\Gamma}) > \frac{1}{2}\right) = 0 \\ \sum_{j=1}^{N_2} j\mathbf{1}\left(P(z_i = j \mid \mathbf{\Gamma}) > \frac{1}{2}\right) & \text{otherwise} \end{cases}$$

i.e. *i* is linked to *j* when the posterior probability of $z_i = j$ is greater than .5

Estimating Link Labels

When estimating $\boldsymbol{z}$, we must account for the within file co-reference structure in MCMC samples



Let $[i]_1$ index the entity represented by record i in $\boldsymbol{X}_1$, and $[j]_2$ index the entity represented by record j in $\boldsymbol{X}_2$. Then the estimated label for entity $[i]_1$ is

$$\widehat{z^*_{[i]_1}} = \begin{cases} [j]_2 & \sum_i P\left(z_i \in \{j \mid j \text{ belongs to } [j]_2\} \mid \boldsymbol{\Gamma}\right) \mathbf{1}(i \in \{i' \mid i' \text{ belongs to } [i]_1)\} > .5 \\ N_2 + \min\{i' \mid i' \text{ belongs to } [i]_1\} & \text{otherwise} \end{cases}$$

Simulating Data with Aliases

Simulating data with aliases

- Simulate 4,000 records
  - with GeCo (Tran et al. 2013)
  - Fields: first name, surname, gender, city (AU), income, age, city (US)
- Impose entity structure:
  - Records with the same first name and US city were assigned to be co-referent alias records
  - US city field removed from data
  - 2,909 entities: 1 alias: 80%; 2 aliases: 12%; 3+ aliases: 8%
- Create 2 corrupted copies of each record

First Name	Surname	Gender	City (AU)	Income	Age	City (US)
peter	lillie-hinrichs	M	perth	648690.01	30	newark
peter	weidenbach	F	canberra	52050.07	33	newark
peter	beams	M	perth	37829.03	33	newark
peter	byers	F	melbourne	1366805.44	29	raleigh
petreece	bishop	M	perth	203509.94	29	new york
petreece	germinario	F	sydney	162218.16	31	newark
peyton	white	M	melbourne	77113.66	31	newark
philip	bacskai	M	melbourne	167950.37	30	newark
philip	pikusa	F	melbourne	1031977.70	26	newark

Simulation Study

Can A-BRL accurately identify links in the presence of multiple alias records?

- Baseline: BRL with additional alias records removed from data.

(Herzog et al. 2007; Abramitzky 2021)

- Compare posterior precision and recall:

Precision:

$$\frac{\text{TP links}}{\text{TP links} + \text{FP links}}$$

Recall:

$$\frac{\text{TP links}}{\text{TP links} + \text{FN links}}$$

Simulation Study

- **3 overlapping pairs of datasets ($n_1 = n_2 = 1000$) created for each (v, p)**
  - Overlap (v): low (5%), medium (25%), high (50%)
  - Expected proportion of retained aliases (p) : high (.75), medium (.5), none (0)
    - 1 alias selected for retention; additional aliases dropped with probability 1- p.
- **Comparison data:** $F = 6$, $K_f = 5$, except f_{gender}, $K_f = 2$.
- **Model Parameters:** $a = 1$, $b = 5$, $\boldsymbol{\alpha}_f = (K_f, \ldots, 1)$, $\boldsymbol{\beta}_f = (1, \ldots, K_f)$

		Expected Proportion of Additional Aliases Retained					
		.75		.5		0 (No Aliases)	
Overlap	**Model**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
5%	BRL	.760	.741	.740	.709	.796	.789
5%	A-BRL	.745	**.804**	**.751**	**.764**	.796	.789
25%	BRL	.894	.825	.884	.828	.885	.829
25%	A-BRL	**.915**	**.911**	**.908**	**.889**	.885	.828
50%	BRL	.913	.852	.920	.854	.919	.853
50%	A-BRL	**.959**	**.928**	**.953**	**.918**	.920	.853

When p = 0, A-BRL and BRL are applied to the same data, and A-BRL = BRL.

Can A-BRL Link These Records?

An excerpt of the records of 2 entities in *Louisiana Kindred* after expanding rows into alias records:

Louisiana Kindred (excerpt)

ID	Name	Relation	Relative	Enslaver	Enslaver Location	…
1	Milly	Mother of	Sarah Ann	William B.G. Taylor	New Orleans, LA	…
1	Milly	Mother of	Sarah Ann	Austin Woolfolk	Augusta, GA	…
1	Milly	Mother of	Hencen	William B.G. Taylor	New Orleans, LA	…
1	Milly	Mother of	Hencen	Austin Woolfolk	Augusta, GA	…
2	Margaret	Mother of	*[unnamed child]*	James Junerarity	New Orleans, LA	…
2	Margaret	Mother of	*[unnamed child]*	John Woolfolk	Augusta, GA	…
⋮	⋮	⋮	⋮	⋮	⋮	

Data Processing Steps

- Expand datasets into alias records
  - Alias information in Relations and Enslavers
- Create comparison data
  - $F = 15$: First Name, Last Name, Gender, Infant, Birth Year, Skin Tone, Kin First Name, Kin Last Name, Kin Type, Event Year, Enslaver First Name, Enslaver Last Name, Enslaver City, Enslaver County, Enslaver State.
- Block data on gender
  - Remove consideration of M/F pairs of F/M pairs.
  - 22,407,192 F/F, F/*NA* and 31,917,672 M/M and M/*NA* comparisons.
- Account for location dependency
  - Keep only the most precise observed comparison of location, set $\omega^f_{ij} = 0$ for the other location comparisons (Resnick 2017).
- Incorporate context of enslaver networks
  - Enslavers co-occurring in data at high rates were considered equivalent (Pritchett 2001; Steckel and Ziebarth 2013).

Fitting A-BRL to Link Records of Enslaved Individuals

- Comparison Fields
  - $f = 1,\ldots,15$
  - $K_f = 5$ for string-valued fields
  - $K_f = 4$ for continuous-valued fields
  - $K_f = 2$ for categorical-valued fields

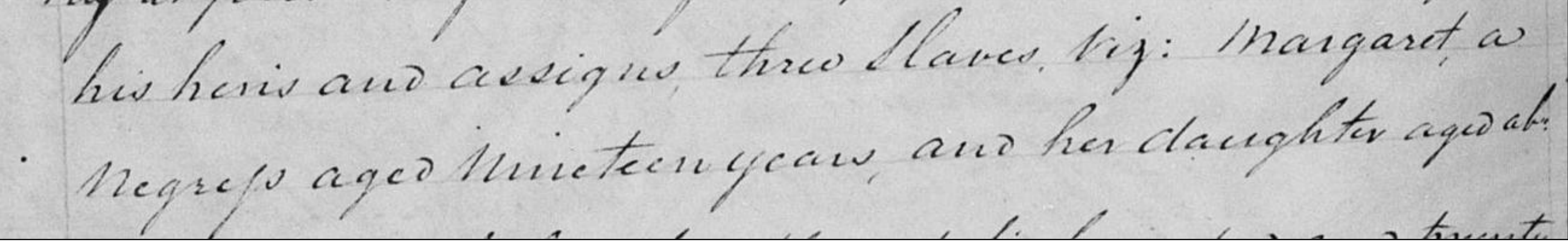
- Agreement-level probabilities
  - $\boldsymbol{\alpha}_f = (K_f, \ldots, 1)$
  - $\boldsymbol{\beta}_f = (1, \ldots, K_f)$

$$\boldsymbol{m}_f = \mathrm{Dir}\left((K_f, \ldots, 1)\right)$$
$$\boldsymbol{u}_f = \mathrm{Dir}\left((1, \ldots, K_f)\right)$$

- Linkage labels
  - $a = 1$
  - $b = 5$

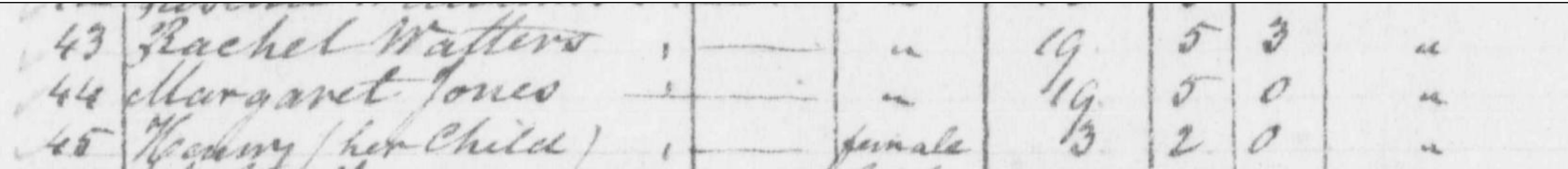
$$P(\boldsymbol{z} \mid L, a, b) \propto \frac{B(a + L, b + N_1 - L)(N_2 - L!)}{B(a, b) N_2!} \mathbf{1}(\boldsymbol{z} \text{ is valid})$$

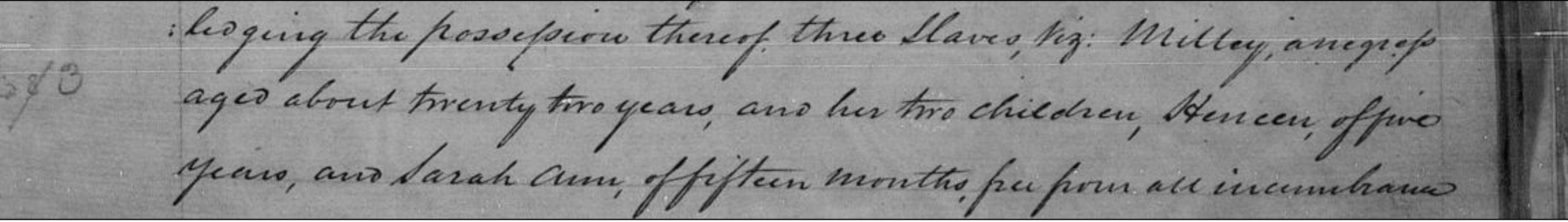
Results



"**Margaret,** a negress aged nineteen years…"

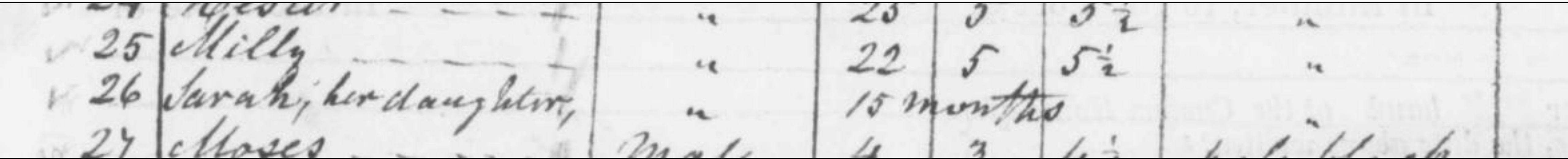
- Margaret in *Louisiana Kindred* was linked to a record for Margaret Jones in *Oceans of Kinfolk*.
- In 1825, she was taken on a brig from Baltimore, MD with 77 other captives to New Orleans, LA.
- She was sold together with her young daughter in New Orleans.
- Her daughter was called Henny.




lodging the possession thereof three Slaves, viz: Milley, a negress
aged about twenty two years, and her two children, Hencen, of five
years, and Sarah Ann, of fifteen months, free from all incumbrance

“**Milly**, a negress aged about twenty two years…”

- Milly in *Louisiana Kindred* was linked to a record of Milly in *Oceans of Kinfolk*.
- In 1826, Milly was taken on a ship from Baltimore, MD with 39 other captives to New Orleans, LA.
- The record of Milly in *Oceans of Kinfolk* had no children listed.
- She was sold together with her young children, Hencen (5) and Sarah Ann (1) in New Orleans in 1826.
- In the ship manifest, an infant named Sarah was listed as Milly’s daughter.
- An apparently unaccompanied child named Moses, aged 4, was also on the manifest.

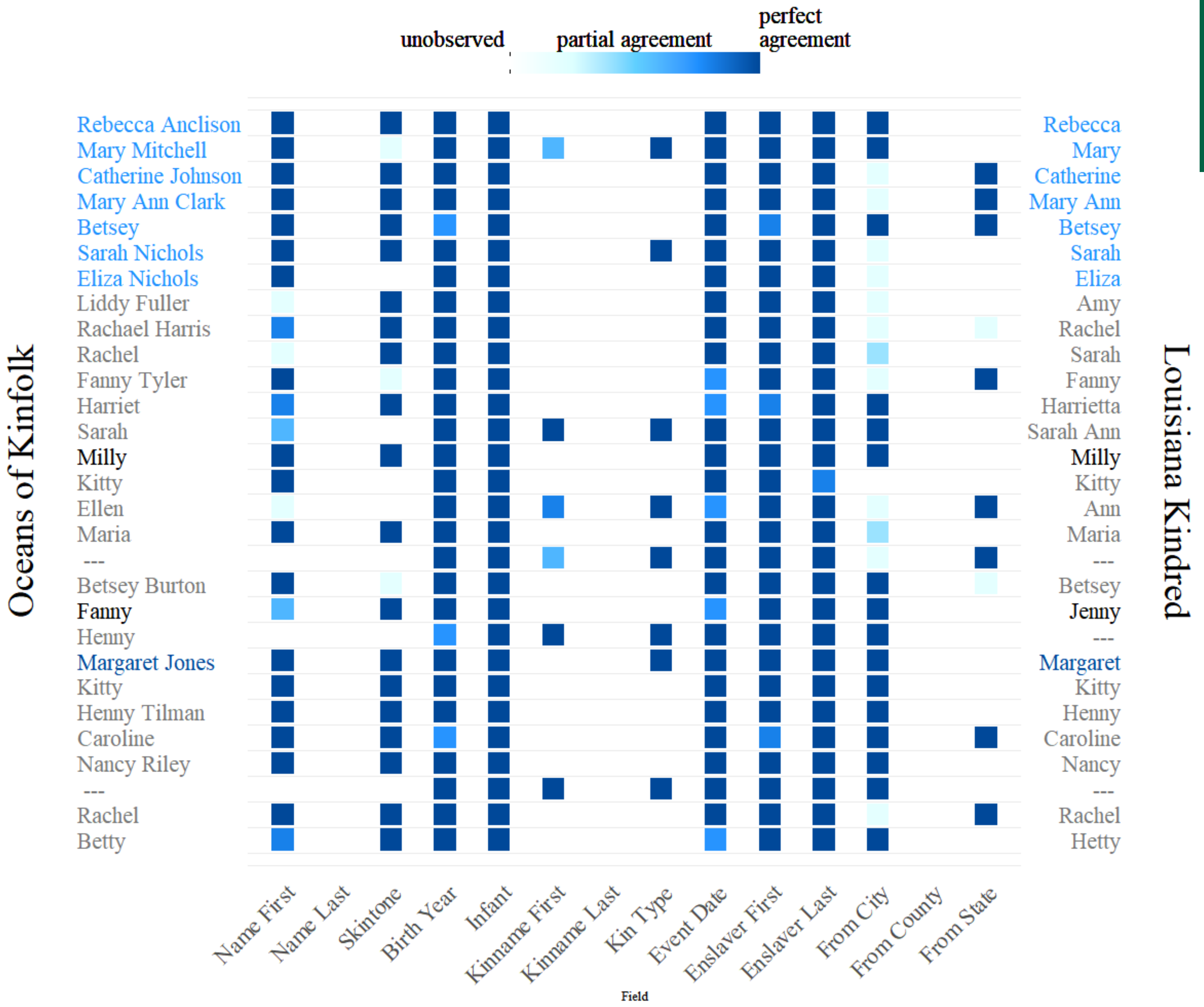


25	Milly	"	22	5	5½	"
26	Sarah, her daughter,	"	15 months			
27	Moses		4	3		

Estimated Links

Links estimated by A-BRL for F/F or F/NA comparisons

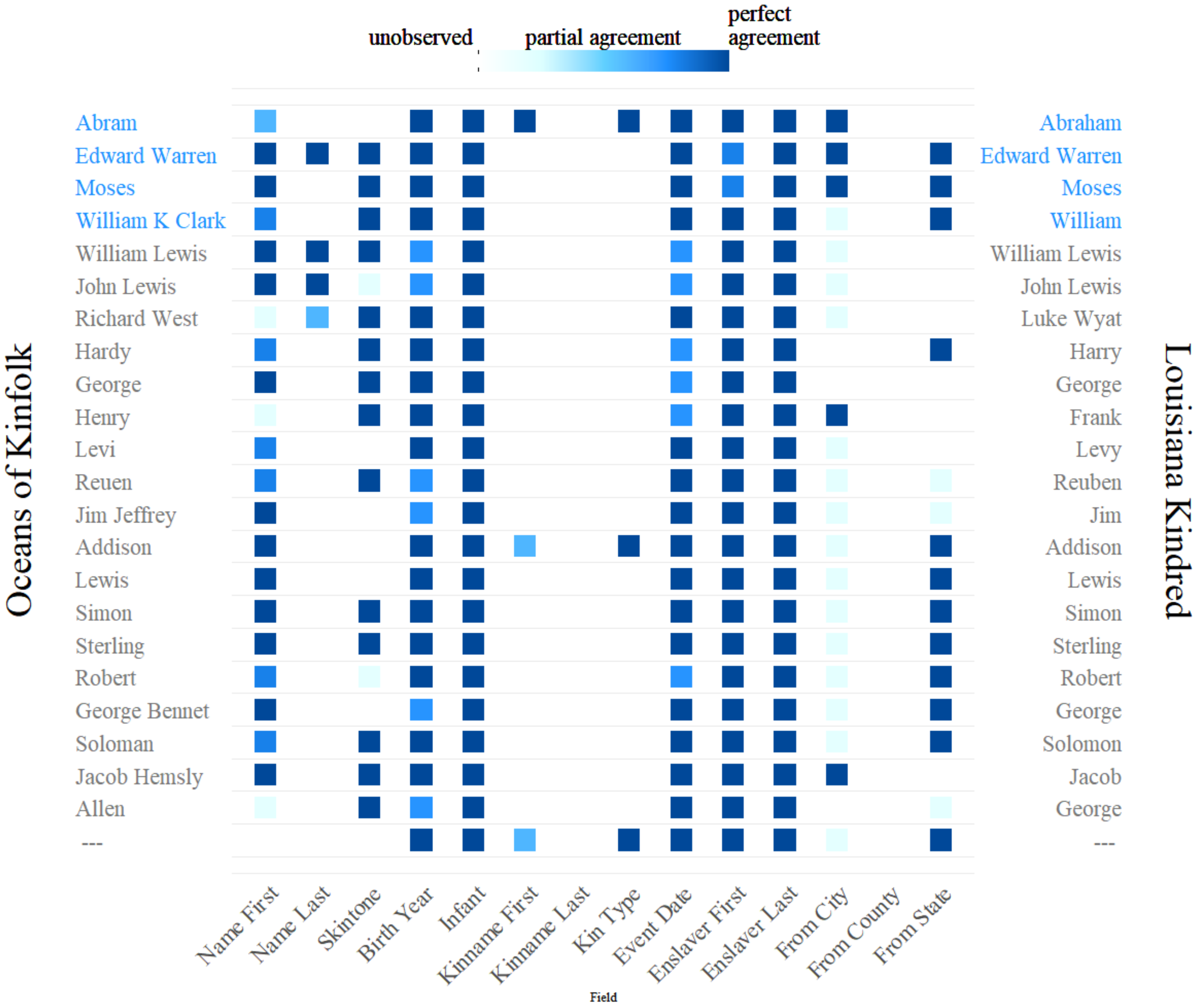
- Names in blue have been validated by historians with manual record linkage.
- Names in gray have not yet been validated.
- Darker squares indicate greater agreement in the corresponding field



Estimated Links

Links estimated by A-BRL for M/M or M/NA comparisons

- Names in blue have been validated by historians with manual record linkage.
- Names in gray have not yet been validated.
- Darker squares indicate greater agreement in the corresponding field



Summary

Contribution:

- Proposed a fully Bayesian record linkage model capable of handling alias records in messy data.

Accomplished:

- Showed competitive performance of A-BRL in simulated scenarios.
- Used A-BRL to perform record linkage on *Oceans of Kinfolk* and *Louisiana Kindred*.

Limitations:

- Have not yet found real data to validate A-BRL
- Simulation study may be too simple to have a complete understanding of how priors affect model behavior
- Certainty in alias structure (See Chapter 4).

Chapter 3

Visualizing Record Linkage Through the Narratives of the Enslaved

Interactive Data Visualization

- Made with Shiny in R (Chang et al. 2024).
  - Shiny is an R package to create applications to allow users to interact with data.
  - Buttons, drop-downs, sliders, etc.

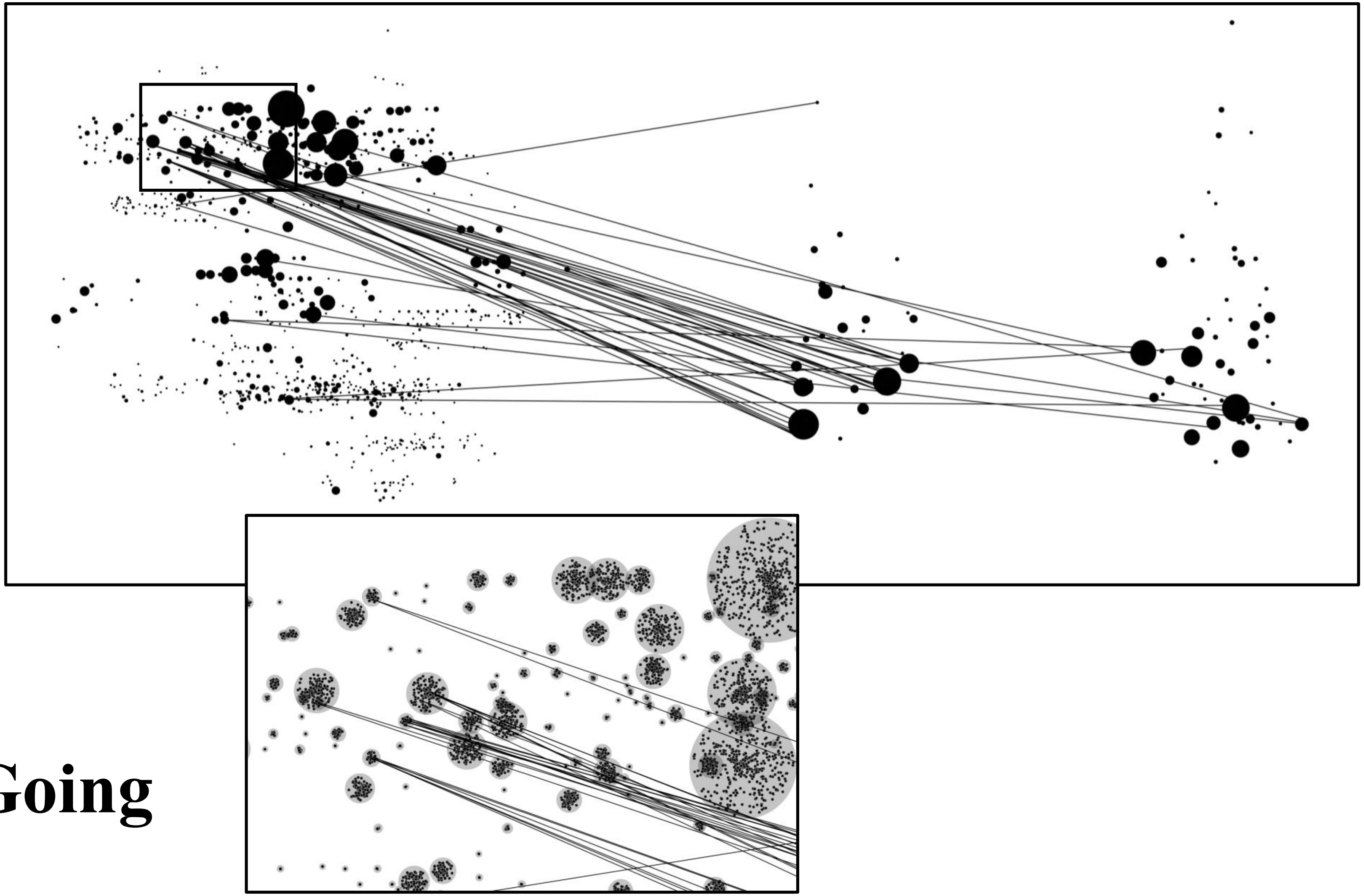
- Uses communication between Shiny elements and Javascript library D3 (Bostock et al. 2011) to create an interactive user experience.
  - Zooming, panning, clicking
  - Animations

- Dissemination options
  - Shinyapps.io
  - R package on CRAN
  - Host on Kinfolkology.com?

Principles and Philosophy

- **Reject Rationality**
  - “Data are not neutral or objective.” (D’Ignazio and Klein 2020)
  - “…Cartesian-based analytical thought that privileges the isolation of certain key aspects of a situation…” (Brasseur, 2003)
- **Invoke Emotion**
  - “Feelings are just as cognitive as other percepts.” (Damasio, 1994)
- **Tell Stories**
  - Provide context
  - Convey knowledge

Demo Application



Where We're Going

Future Work

Contribution:

- Proposed a fully Bayesian record linkage model capable of handling alias records in messy data.

Accomplished:

- Showed competitive performance of A-BRL in simulated scenarios.
- Used A-BRL to perform record linkage on *Oceans of Kinfolk* and *Louisiana Kindred*.

Future Work:

- Finish interactive visualization and make available for public use.
- Further validation of A-BRL performance.
  - Additional simulation studies.
  - Real data for validation.
- Consider probabilistic estimates of internal alias structures to link data with both known and unknown internal links (Dissertation Chapter 4).

References

***[Untitled Slide]* (3)**

"United States, Census, 1880", FamilySearch (https://www.familysearch.org/ark:/61903/1:1:M4NB-K1V : Tue Oct 21 21:19:03 UTC 2025), Entry for Robt Taylor and Matilda Taylor, 1880.

"United States, Census, 1870", FamilySearch (https://www.familysearch.org/ark:/61903/1:1:MHK4-3TG : Fri Jan 17 00:30:18 UTC 2025), Entry for Robert Slaton and Matilda Slaton, 1870.

"United States, Census, 1860", FamilySearch (https://www.familysearch.org/ark:/61903/1:1:MHD4-J8Z : Mon Jul 08 02:50:27 UTC 2024), Entry for N Slaton and Franklin Smith, 1860.

"United States, Census (Slave Schedule), 1860", FamilySearch (https://www.familysearch.org/ark:/61903/1:1:WKNN-JQPZ : Sat Mar 09 15:49:17 UTC 2024), Entry for N Slaton and , 1860.

Applications of Record Linkage (7)

Jaro, Matthew A. “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” *Journal of the American Statistical Association* 84, no. 406 (1989): 414–20, https://doi.org/10.1080/01621459.1989.10478785.

Winkler, William E. and Yves Thibaudeau, *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census* (US Bureau of the Census Washington, DC, 1991), https://www.academia.edu/download/34942646/rr91-9.pdf.

Winkler, William E. *Record Linkage Software and Methods for Merging Administrative Lists* (US Bureau of the Census, 2001), https://ecommons.cornell.edu/bitstreams/4f0722d7-5105-4f8b-a1dc-6432b1c23f61/download.

Kaplan, Andee et al., “A Practical Approach to Proper Inference with Linked Data,” *The American Statistician* 76, no. 4 (2022): 384–93, https://doi.org/10.1080/00031305.2022.2041482.

Sosa, Juan and Rodríguez, Abel .“A Bayesian Record Linkage Model Incorporating Relational Data,” *Applied Stochastic Models in Business and Industry* 39, no. 6 (2023): 755–71, https://doi.org/10.1002/asmb.2792.

L. Drew et al., “A Bayesian Record Linkage Approach to Applications in Tree Demography Using Overlapping LiDAR Scans,” arXiv:2501.13285, preprint, arXiv, June 5, 2025, https://doi.org/10.48550/arXiv.2501.13285.

Xinyi Lu et al., “Improving Wildlife Population Inference Using Aerial Imagery and Entity Resolution,” *Journal of Agricultural, Biological and Environmental Statistics* 27, no. 2 (2022): 364–81, https://doi.org/10.1007/s13253-021-00484-w.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59 (3): 865–918

Richard H. Steckel and Nicolas Ziebarth, “A Troublesome Statistic: Traders and Coastal Shipments in the Westward Movement of Slaves,” *The Journal of Economic History* 73, no. 3 (2013): 792–809, https://doi.org/10.1017/S0022050713000612.

Mauricio Sadinle, “Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach,” *The Annals of Applied Statistics* 8, no. 4 (2014): 2404–34, https://doi.org/10.1214/14-AOAS779.

Maria Gargiulo et al., *Deaths in Custody during the Armed Conflict in Syria, 2011–2023* (Human Rights Data Analysis Group, 2024).

Why This Data? (11)

Jonathan B. Pritchett, “Quantitative Estimates of the United States Interregional Slave Trade, 1820-1860,” *The Journal of Economic History* 61, no. 2 (2001): 467–75.

Richard H. Steckel and Nicolas Ziebarth, “A Troublesome Statistic: Traders and Coastal Shipments in the Westward Movement of Slaves,” *The Journal of Economic History* 73, no. 3 (2013): 792–809, https://doi.org/10.1017/S0022050713000612.

Jennie K. Williams, “Trouble the Water: The Baltimore to New Orleans Coastwise Slave Trade, 1820–1860.,” *Slavery & Abolition* 41, no. 2 (2020): 275–303, 143545159, https://doi.org/10.1080/0144039X.2019.1660509.

References

William D. Jones, “Beyond New Orleans: Forced Migrations To, From, and In Louisiana, 1820?1860,” *Louisiana History: The Journal of the Louisiana Historical Association* 62, no. 4 (2021): 429–68.

Data on Enslaved People (12)

David Thomas and Simon Fowler, *The Silence of the Archive* (Facet, 2017), https://doi.org/10.29085/9781783301577.

Probabilistic Record Linkage (13)

Ivan P. Fellegi and Alan B. Sunter, “A Theory for Record Linkage,” *Journal of the American Statistical Association* 64, no. 328 (1969): 1183–210, https://doi.org/10.2307/2286061.

Mauricio Sadinle, “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association* 112, no. 518 (2017): 600–612, https://doi.org/10.1080/01621459.2016.1148612.

Andrea Tancredi and Brunero Liseo, “A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems1,” *The Annals of Applied Statistics* 5, no. 2B (2011): 1553–85.

Rebecca C. Steorts et al., “A Bayesian Approach to Graphical Record Linkage and Deduplication,” *Journal of the American Statistical Association* 111, no. 516 (2016): 1660–72, https://doi.org/10.1080/01621459.2015.1105807.

Marco Fortini et al., “On Bayesian Record Linkage,” *Research in Official Statistics* 4 (2001).

Comparison Data for Classical RL (14)

Ivan P. Fellegi and Alan B. Sunter, “A Theory for Record Linkage,” *Journal of the American Statistical Association* 64, no. 328 (1969): 1183–210, https://doi.org/10.2307/2286061.

Bayesian Record Linkage (20)

Marco Fortini et al., “On Bayesian Record Linkage,” *Research in Official Statistics* 4 (2001).

Andrea Tancredi and Brunero Liseo, “A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems1,” *The Annals of Applied Statistics* 5, no. 2B (2011): 1553–85.

Rebecca C. Steorts et al., “A Bayesian Approach to Graphical Record Linkage and Deduplication,” *Journal of the American Statistical Association* 111, no. 516 (2016): 1660–72, https://doi.org/10.1080/01621459.2015.1105807.

Mauricio Sadinle, “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association* 112, no. 518 (2017): 600–612, https://doi.org/10.1080/01621459.2016.1148612.

Beta-Bipartite Prior (21)

Mauricio Sadinle, “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association* 112, no. 518 (2017): 600–612, https://doi.org/10.1080/01621459.2016.1148612.

Alias Records (23)

Hadley Wickham, “Tidy Data,” *Journal of Statistical Software* 59 (September 2014): 1–23, https://doi.org/10.18637/jss.v059.i10.

Aliased Beta-Bipartite Prior (BRL) (27)

Mauricio Sadinle, “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association* 112, no. 518 (2017): 600–612, https://doi.org/10.1080/01621459.2016.1148612.

Estimating Link Labels (28)

Mauricio Sadinle, “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association* 112, no. 518 (2017): 600–612, https://doi.org/10.1080/01621459.2016.1148612.

References

Comparison Data for Classical RL (14)

Ivan P. Fellegi and Alan B. Sunter, “A Theory for Record Linkage,” *Journal of the American Statistical Association* 64, no. 328 (1969): 1183–210, https://doi.org/10.2307/2286061.

Simulating Data with Aliases (30)

Khoi-Nguyen Tran et al., “GeCo: An Online Personal Data Generator and Corruptor,” *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (New York, NY, USA), CIKM ’13, Association for Computing Machinery, October 27, 2013, 2473–76, https://doi.org/10.1145/2505515.2508207.

Simulation Study (31)

Thomas N. Herzog et al., *Data Quality and Record Linkage Techniques* (Springer, 2007), https://doi.org/10.1007/0-387-69505-2.

Ran Abramitzky et al., “Automated Linking of Historical Data,” *Journal of Economic Literature* 59, no. 3 (2021): 865–918, https://doi.org/10.1257/jel.20201599.

Interactive Data Visualization (43)

Winston Chang et al., *Shiny: Web Application Framework for R* (2024), https://CRAN.R-project.org/package=shiny.

Michael Bostock et al., “D3: Data-Driven Documents,” *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011, http://vis.stanford.edu/papers/d3.

Principles and Philosophy (44)

Bayesian Record Linkage (20) Catherine D’Ignazio and Lauren F. Klein, *Data Feminism* (The MIT Press, 2020), https://doi.org/10.7551/mitpress/11805.001.0001.

Lee E. Brasseur, *Visualizing Technical Information: A Cultural Critique*, ed. Charles H. Sides, Baywood’s Technical Communications Series (Baywood Publishing Company, Inc., 2003).

Antonio R. Damasio, *Descartes’ Error: Emotion, Reason, and the Human Brain* (Avon Books, 1994).

Chapter 4: Proposal

Probabilistic De-Duplication for Alias Identification To Apply Bayesian Bipartite Record Linkage to Crowd-Sourced Data.

- Develop/Specify process
  - 2-step process? Estimate aliases, then perform A-BRL with estimated alias structure
  - Joint Model? Perform alias estimation as part of fully Bayesian record linkage model

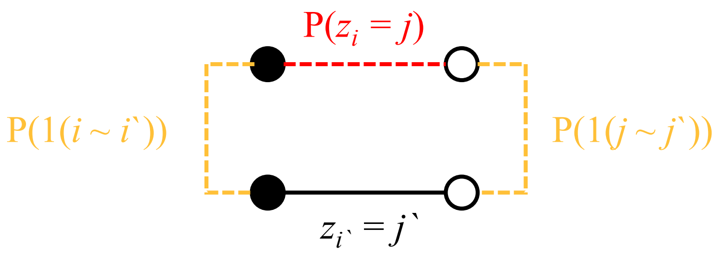
- Perform simulation studies
  - Compare model to existing record linkage models?
  - Compare model to A-BRL?
  - Compare deduplication with existing methods of deduplication?

- Apply model to link crowd-sourced data
  - Freedom on the Move
  - The Underground Railroad (William Still)

Chapter 4: Estimating Links with Uncertain Aliases

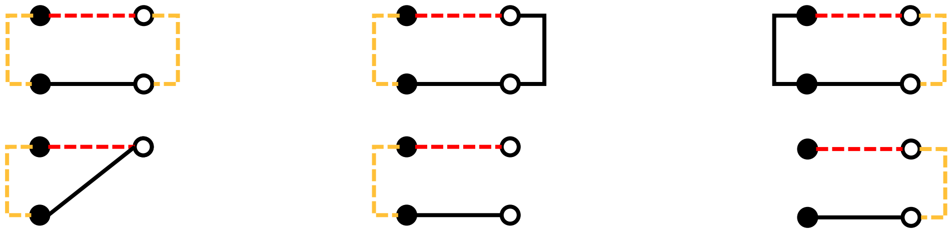
A-BRL Prior: $P(z_i = j \mid \mathbf{1}(i \sim i'), \mathbf{1}(j \sim j'), z_{i'} = j')$

- > 0 only if $1(i \sim i') = 0$, $1(j \sim j') = 0$
- $1(i \sim i') = 0$, $1(j \sim j') = 0$ assumed known



- How to estimate links when $\mathbf{1}(i \sim i')$, $\mathbf{1}(j \sim j')$ random?
  - Use weighting:

  $P(z_i = j, z_{i'} = j' \mid 1(i \sim i') = 0, 1(j \sim j') = 0) P(1(i \sim i') = 0)\ P(1(j \sim j') = 0)$
  - Iteratively update alias structure



- How to handle/determine transitive conflicts?
  - For weighting: Threshold?

  Multiply prob. by $\mathbf{1}_{\text{thr}}(P(1(i \sim i') = 0) < p_1, P(1(j \sim j') = 0) < p_2)$
  - Iterative updating: Not sure yet

Chapter 4: Simulation Studies

Can a deduplication model even perform alias-identification?

- Consider fundamental differences in duplicates vs. aliases
  - Alias differences are non-random
  - Aliases are likely to disagree in certain fields
- Test performance of deduplication models on data with erroneous duplicates vs data with a latent/pseudo-latent alias structure.

How should we gauge the performance of our approach to linking with alias uncertainty?

- Compare within the scope of A-BRL?
- Compare to other models of record linkage that don’t consider alias structure at all?

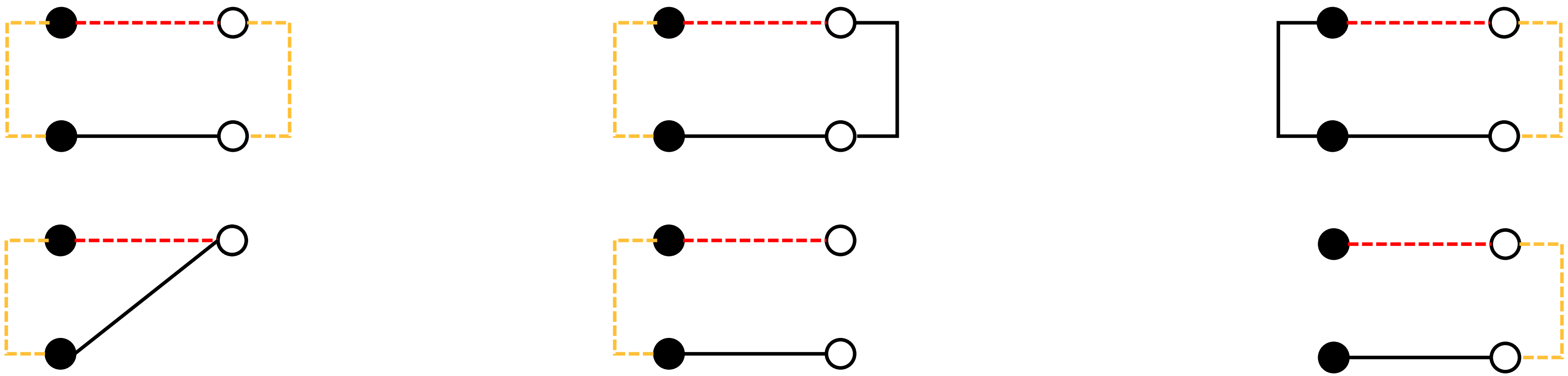
Data will need to be generated to perform proposed simulation study.

- Use GeCo (Tran et al. 2013)

Chapter 4

Data

- William Still's The Underground Railroad
  - Contains known aliases
- Freedom on The Move
  - Crowd-sourced
  - Suspect alias records, but unknown co-reference structure
  - Apply de-duplication to estimate alias structure
  - Use posterior probabilities as weights in place of alias matrix.



$\mathbf{1}(i \sim i`)$, $\mathbf{1}(j \sim j`)$ fixed:

$P(z_i = j \mid \mathbf{1}(i \sim i`), \mathbf{1}(j \sim j`), z_{i`} = j`)$

- $j` = j$: 0
- $j` \neq j$:
  - 1(j ~ j`) = 1: 0
  - 1(j ~ j`) = 0:
    - 1(i ~ i`) = 1: 0
    - 1(i ~ i`) = 0: P(zi = j, zi` = j`)

$\mathbf{1}(i \sim i`)$, $\mathbf{1}(j \sim j`)$ random:

- $P(z_i = j \mid \mathbf{1}(i \sim i`), \mathbf{1}(j \sim j`), z_{i`} = j`)$

$\alpha\ P(z_i = j, z_{i`} = j` \mid \mathbf{1}(i \sim i`) = 0, \mathbf{1}(j \sim j`) = 0) P(\mathbf{1}(i \sim i`) = 0)\ P(\mathbf{1}(j \sim j`) = 0)$