End-to-End Bayesian Entity Resolution

Rebecca C. Steorts

Department of Statistical Science, affiliated faculty in
Computer Science, Biostatistics and Bioinformatics, the
information initiative at Duke (iiD) and
the Social Science Research Institute (SSRI)
Duke University and U.S. Census Bureau

This work is supported by NSF CAREER Award 1652431.

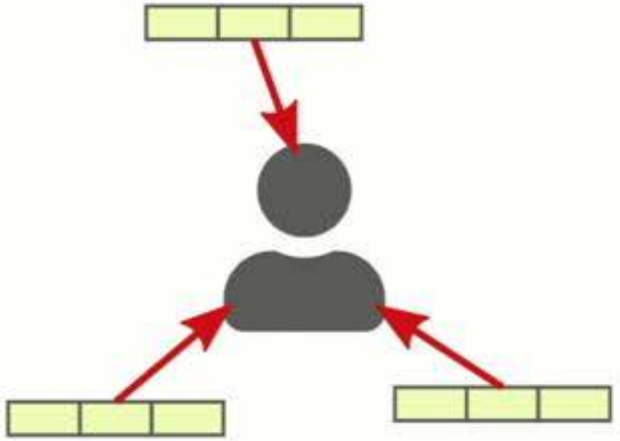
September 6, 2019

Entity resolution

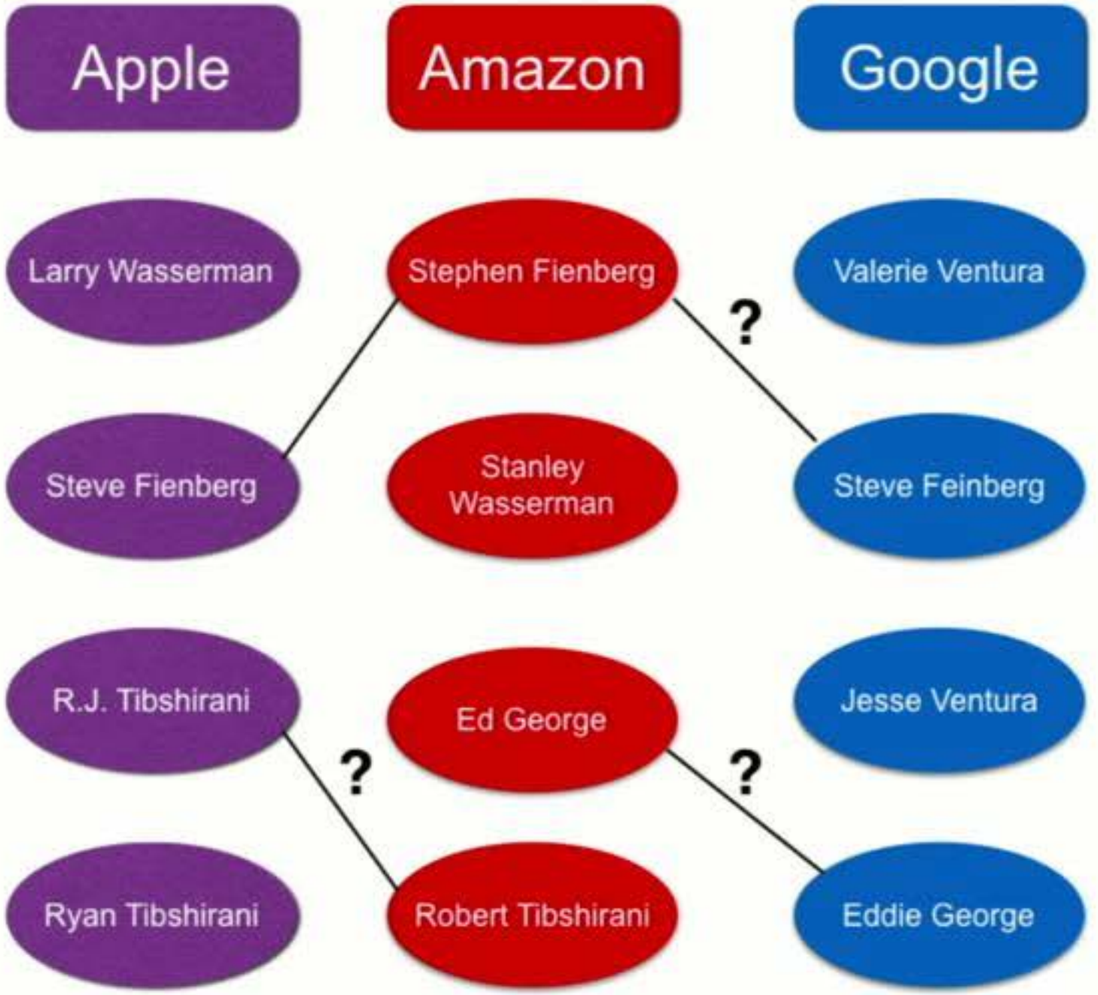
Identifying records across and/or within data sources that refer to the same entities

Also known as:

- record linkage
- data matching
- de-duplication
- data integration



The entity resolution graph

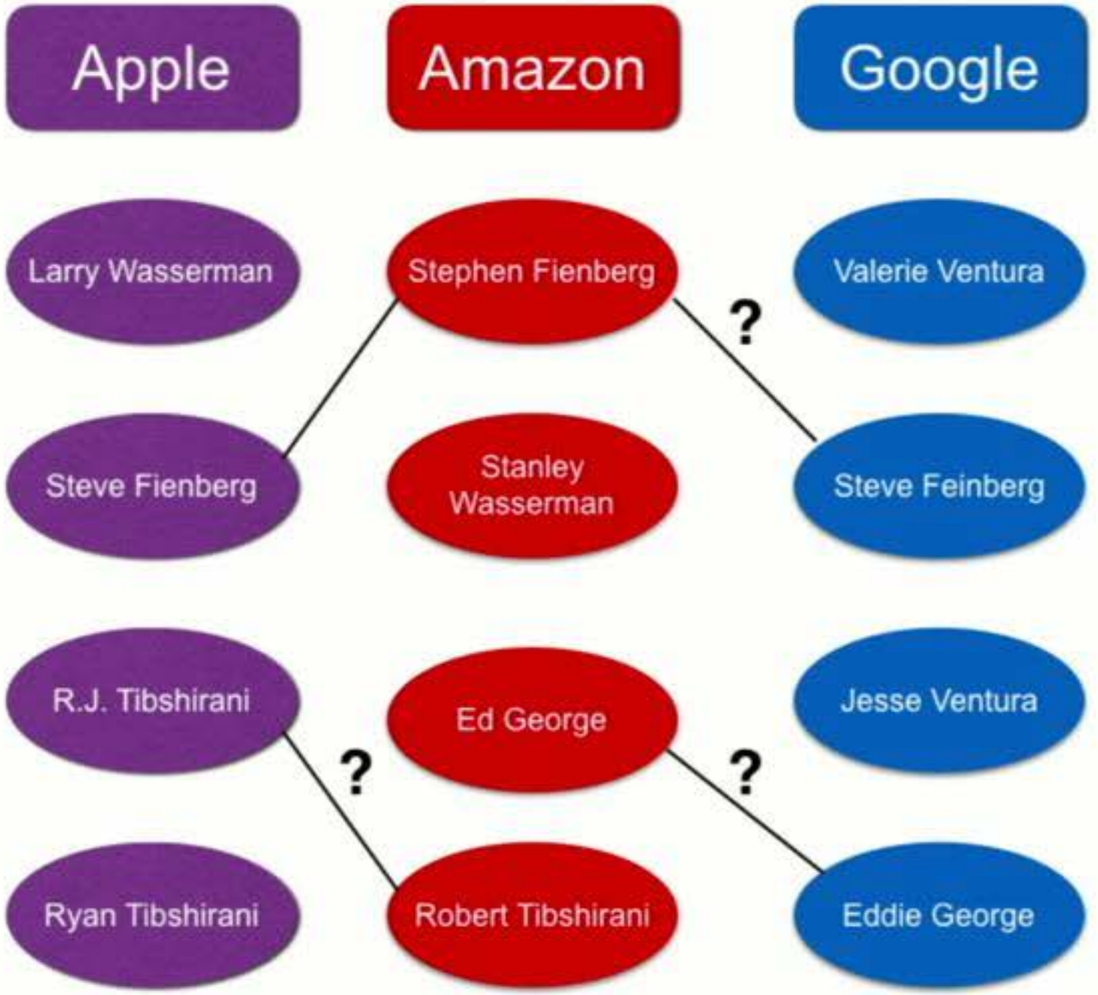


The node of Larry Wasserman

The node of Larry Wasserman

1014 Murray Hill Avenue
Pittsburgh, PA 15217
412-361-3146

The entity resolution graph



240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

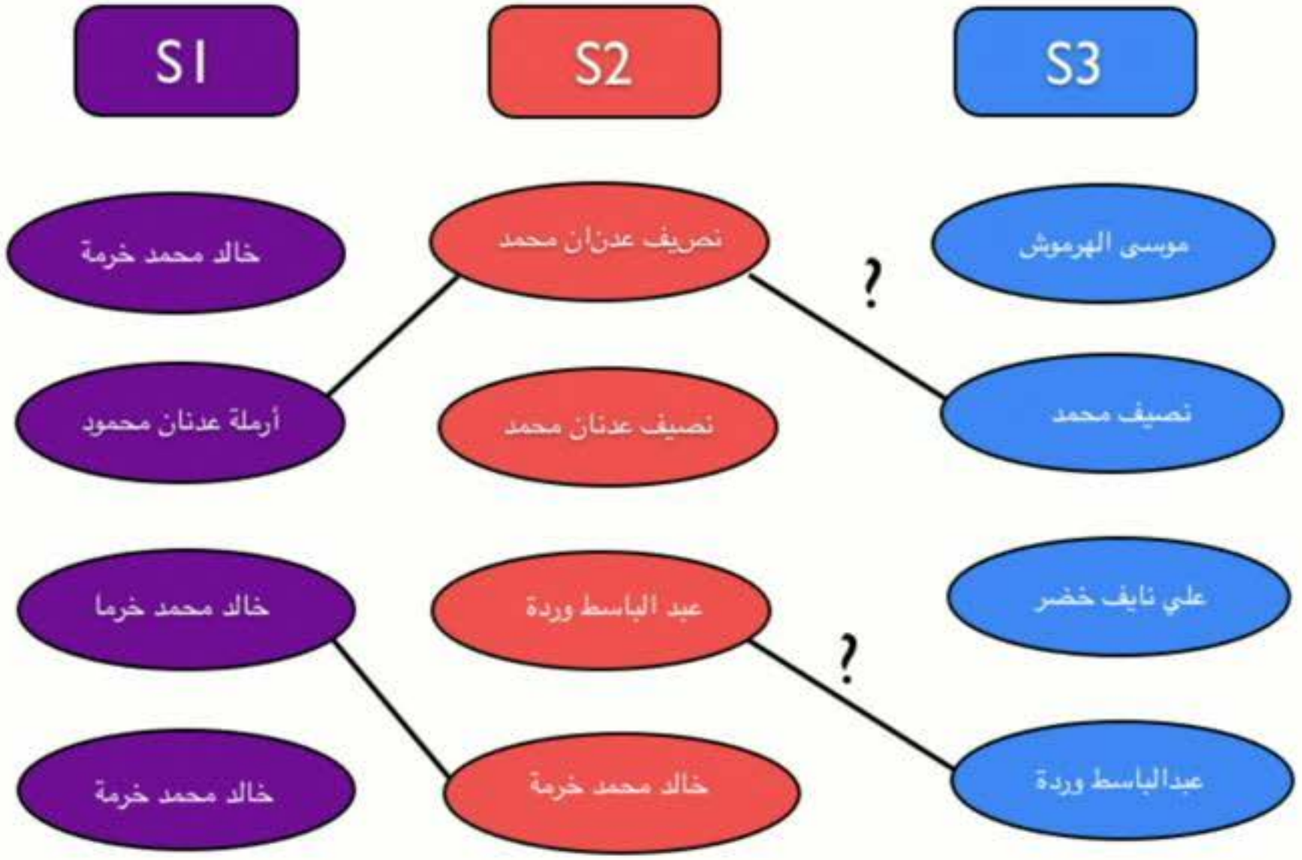
537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

These are clearly not the *same* Steve Fienberg!

Syrian Civil War



Entity Resolution

Why is entity resolution difficult?

Goals of Entity Resolution

Suppose that we have a total of N records in k databases.

1. We seek models that are much less than $O(N^k)$.
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.
3. We seek models and algorithms to handle unbalanced data (containing duplications).

Established ER methods

Common unsupervised methods in statistics:

1. probabilistic linking (Fellegi-Sunter theory)
2. deterministic linking

Established ER methods

Common unsupervised methods in statistics:

1. probabilistic linking (Fellegi-Sunter theory)
2. deterministic linking

Drawbacks:

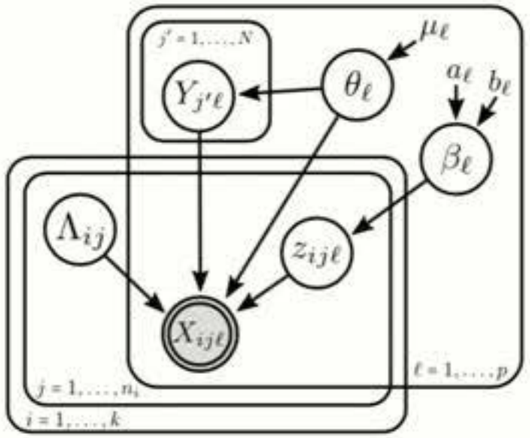
- pairs of records are assessed independently
- awkward post-processing step (transitive closure)
- subjectivity in setting the decision threshold
- lack of uncertainty quantification
- scalability achieved through deterministic blocking

Recent Bayesian methods

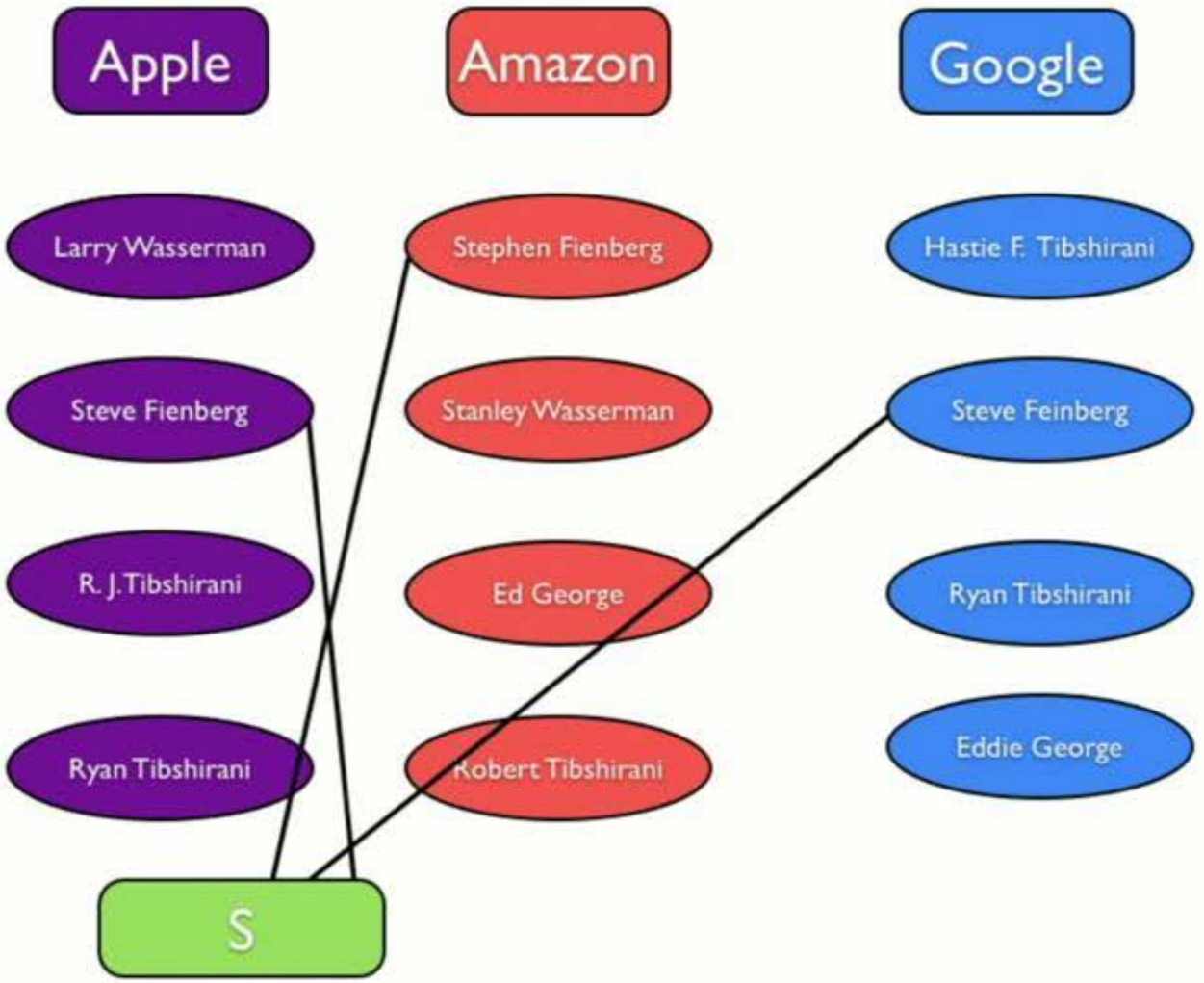
- Dealing with big data means merging large, noisy databases.
  - Such databases have severe amounts of noise.
- Entity resolution requires sophisticated graph structures. [Gutman et. al (2013)].
  - One approach is to use a bipartite graph for latent entities.
  - Never link records to records.

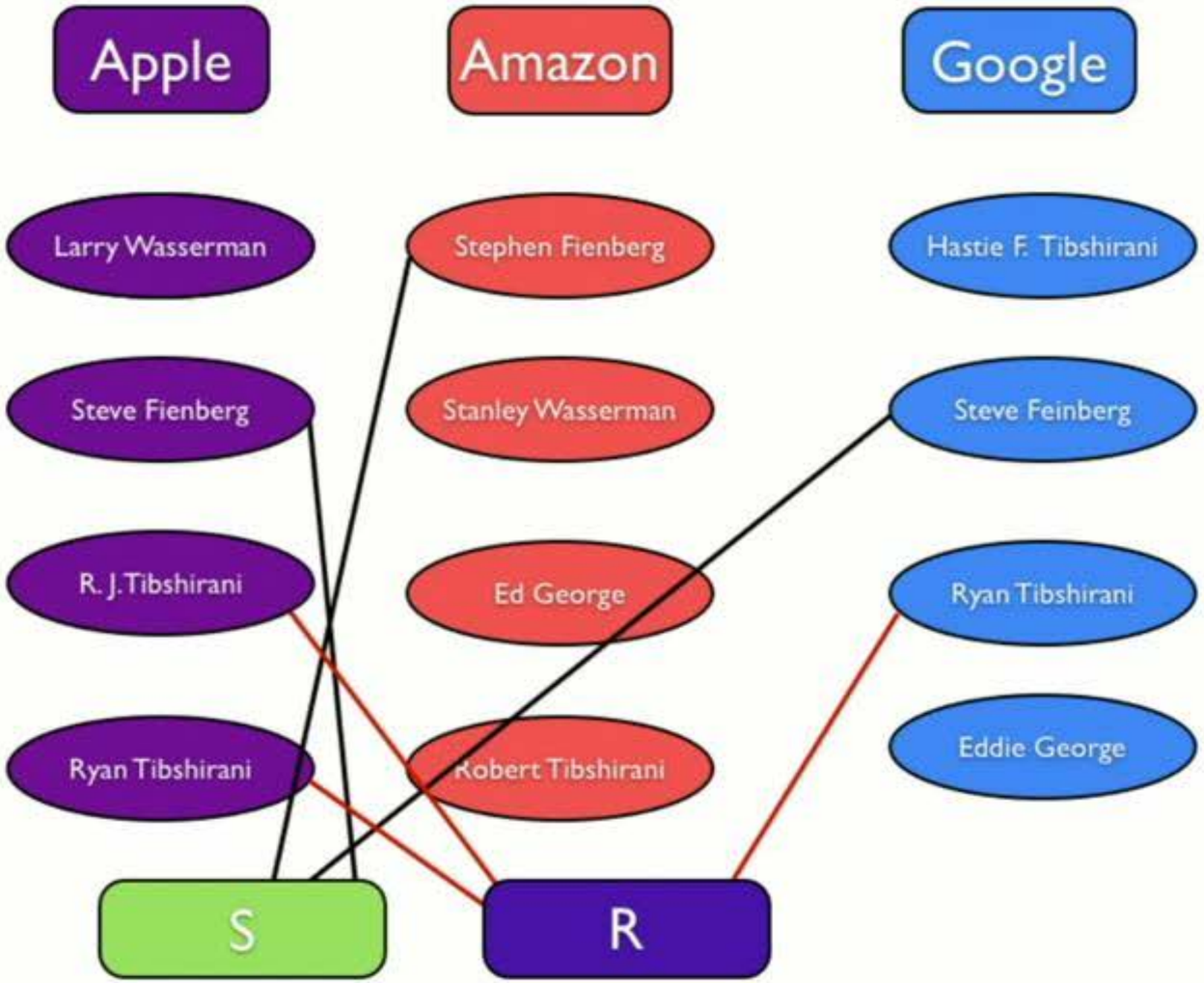
Recent Bayesian methods

- Dealing with big data means merging large, noisy databases.
  - Such databases have severe amounts of noise.
- Entity resolution requires sophisticated graph structures. [Gutman et. al (2013)].
  - One approach is to use a bipartite graph for latent entities.
  - Never link records to records.
- Computational speed-ups: eliminate low-probability matches.



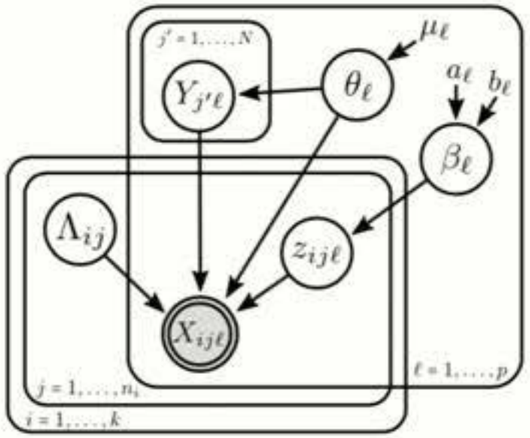
[Tancredi and Liseo (2011), **RCS**, Hall, Fienberg (2014, 2016); Sadinle (2014, 2016), **RCS** (2015), **RCS** et al. (2017), (2018)].


Apple
Amazon
Google
Larry Wasserman
Steve Fienberg
R. J. Tibshirani
Ryan Tibshirani
Stephen Fienberg
Stanley Wasserman
Ed George
Robert Tibshirani
Hastie F. Tibshirani
Steve Feinberg
Ryan Tibshirani
Eddie George
S

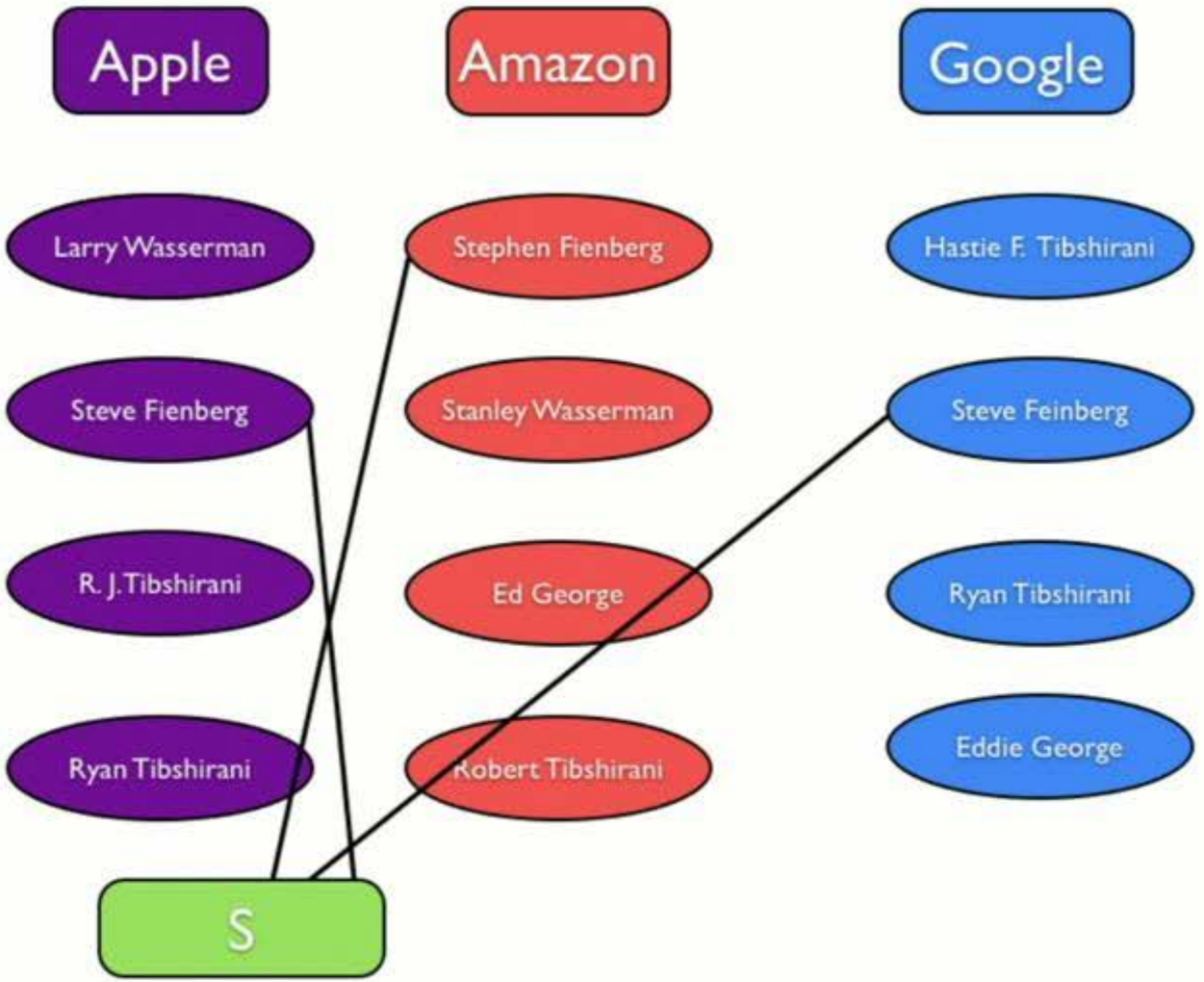

Apple
Amazon
Google
Larry Wasserman
Steve Fienberg
R. J. Tibshirani
Ryan Tibshirani
Stephen Fienberg
Stanley Wasserman
Ed George
Robert Tibshirani
Hastie F. Tibshirani
Steve Feinberg
Ryan Tibshirani
Eddie George
S
R

Recent Bayesian methods

- Dealing with big data means merging large, noisy databases.
  - Such databases have severe amounts of noise.
- Entity resolution requires sophisticated graph structures. [Gutman et. al (2013)].
  - One approach is to use a bipartite graph for latent entities.
  - Never link records to records.
- Computational speed-ups: eliminate low-probability matches.

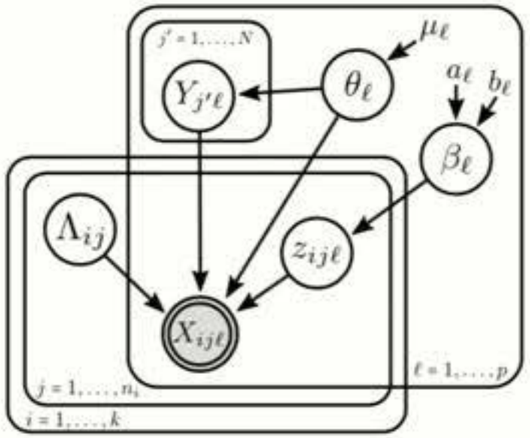


[Tancredi and Liseo (2011), **RCS**, Hall, Fienberg (2014, 2016); Sadinle (2014, 2016), **RCS** (2015), **RCS** et al. (2017), (2018)].

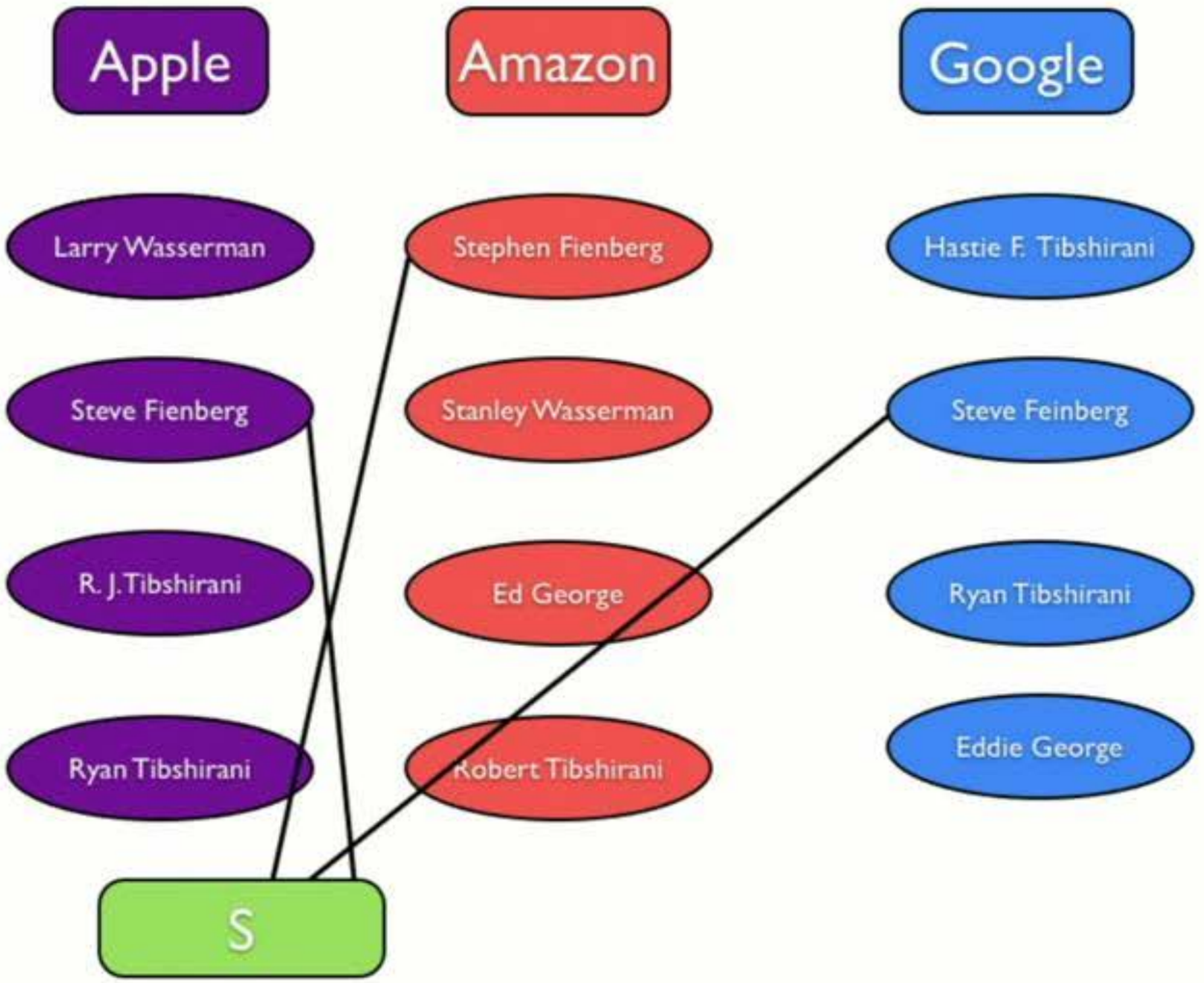

Apple
Amazon
Google
Larry Wasserman
Steve Fienberg
R. J. Tibshirani
Ryan Tibshirani
Stephen Fienberg
Stanley Wasserman
Ed George
Robert Tibshirani
Hastie F. Tibshirani
Steve Feinberg
Ryan Tibshirani
Eddie George
S

Recent Bayesian methods

- Dealing with big data means merging large, noisy databases.
  - Such databases have severe amounts of noise.
- Entity resolution requires sophisticated graph structures. [Gutman et. al (2013)].
  - One approach is to use a bipartite graph for latent entities.
  - Never link records to records.
- Computational speed-ups: eliminate low-probability matches.



[Tancredi and Liseo (2011), **RCS**, Hall, Fienberg (2014, 2016); Sadinle (2014, 2016), **RCS** (2015), **RCS** et al. (2017), (2018)].


Apple
Amazon
Google
Larry Wasserman
Steve Fienberg
R. J. Tibshirani
Ryan Tibshirani
Stephen Fienberg
Stanley Wasserman
Ed George
Robert Tibshirani
Hastie F. Tibshirani
Steve Feinberg
Ryan Tibshirani
Eddie George
S

Our Goal

Scaling Bayesian ER methods to millions of records without sacrificing accuracy and crucially giving uncertainty of the ER task

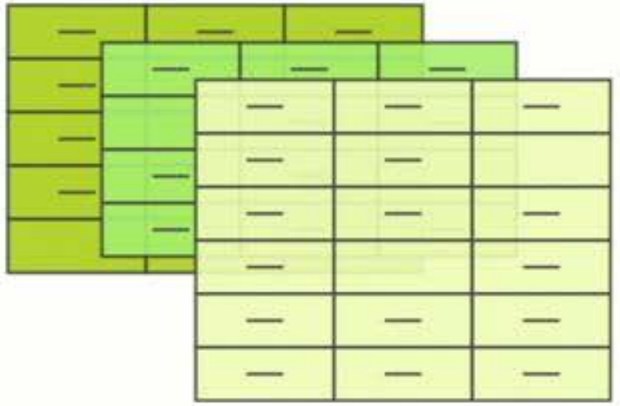
1. Incorporating auxiliary partitions that induce conditional independencies between the entities. This enables distributed inference at the partition-level, while crucially preserving the marginal posterior of the original model.
2. A partition function (responsible for partitioning the entities), which groups similar entities together while achieving well-balanced partitions.
3. Application of partially-collapsed Gibbs sampling in the context of distributed computing.
4. Improving computational efficiency:
   a) Sub-quadratic algorithm for updating links based on indexing.
   b) Truncation of the attribute similarities.
   c) Perturbation sampling algorithm for updating the entity attributes, which relies on the Vose-Alias method.

Marchant, **RCS**, Kaplan, Rubinstein, and Elazar (2019).

Problem setup

Key assumptions:

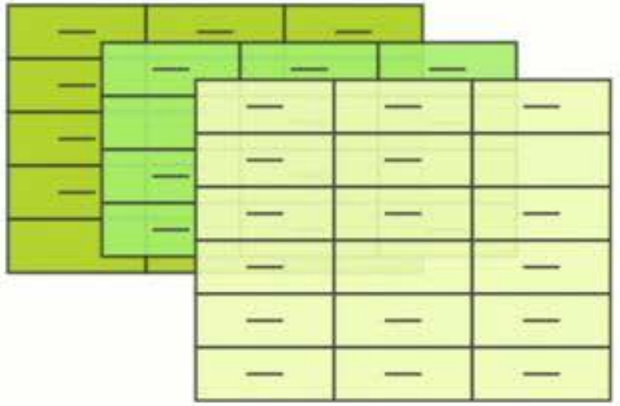
- multiple tables/sources
- duplicates within and across tables
- attributes are aligned
- attributes are discrete
- some missing values
- no ground truth (unsupervised)



Problem setup

Key assumptions:

- multiple tables/sources
- duplicates within and across tables
- attributes are aligned
- attributes are discrete
- some missing values
- no ground truth (unsupervised)



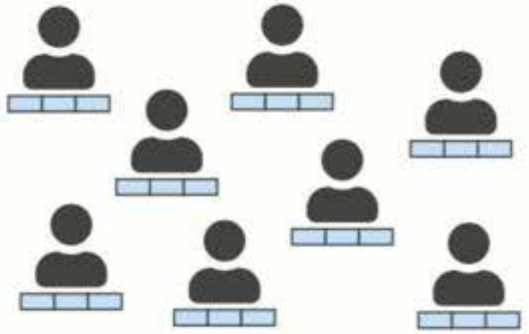
Output: approximate posterior distribution over the linkage structure

`blink` : Bayesian linkage

Latent Entities

- Fixed population of entities $\mathcal{E} = \{1, \ldots, E\}$
- Entity attribute a has a finite domain with distribution ϕ_a
- Generate the value for attribute a of entity e by

$$y_{ea} \sim \text{Discrete}(\phi_a)$$

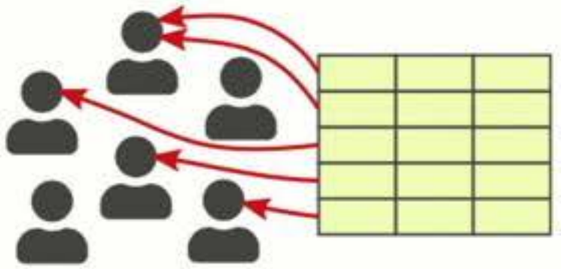


`blink` : Bayesian linkage

Linkage structure

- Record r in table t is linked to an entity $\lambda_{tr} \in \mathcal{E}$ uniformly at random, viz.

$$\lambda_{tr} \sim \text{DiscreteUniform}(\mathcal{E})$$



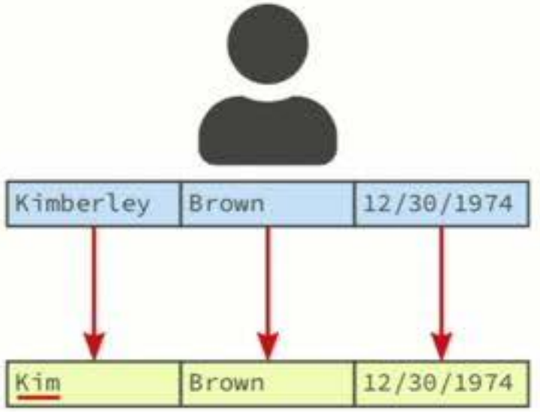
`blink` : Bayesian linkage

- A distortion probability is associated with each attribute a and table t

$$\theta_{ta} \sim \text{Beta}(\alpha_a, \beta_a)$$

- The value for attribute a of record r in table t follows a hit-miss model

$$\begin{aligned} z_{tra}|\theta_{ta} &\sim \text{Bernoulli}(\theta_{ta}) \\ x_{tra}|z_{tra}, y_{\lambda_{tr}a} &\sim (1 - z_{tra})\delta(y_{\lambda_{tr}a}) \\ &\quad + z_{tra}\,\text{Discrete}(\psi_a[y_{\lambda_{tr}a}]) \end{aligned}$$



- Distortion distribution depends on similarity function $\text{sim}_a(\cdot,\cdot)$:

$$\psi_a[y](x) \propto \phi_a(x)\exp\left(\text{sim}_a(y,x)\right)$$

d-blink: a more general and scalable model

We propose a distributed version of blink called d-blink
↪ distributed, more general model, faster inference algorithms

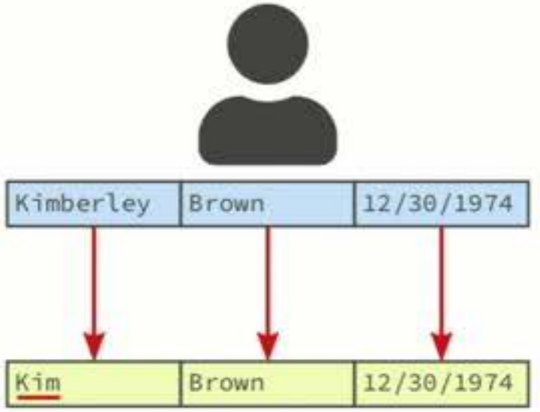
`blink` : Bayesian linkage

- A distortion probability is associated with each attribute a and table t

$$\theta_{ta} \sim \text{Beta}(\alpha_a, \beta_a)$$

- The value for attribute a of record r in table t follows a hit-miss model

$$\begin{aligned} z_{tra} | \theta_{ta} &\sim \text{Bernoulli}(\theta_{ta}) \\ x_{tra} | z_{tra}, y_{\lambda_{tr} a} &\sim (1 - z_{tra}) \delta(y_{\lambda_{tr} a}) \\ &\quad + z_{tra} \, \text{Discrete}(\psi_a[y_{\lambda_{tr} a}]) \end{aligned}$$



- Distortion distribution depends on similarity function $\text{sim}_a(\cdot, \cdot)$:

$$\psi_a[y](x) \propto \phi_a(x) \exp\left(\text{sim}_a(y, x)\right)$$

d-blink: a more general and scalable model

We propose a distributed version of blink called d-blink
↪ distributed, more general model, faster inference algorithms

Modeling differences in d-blink

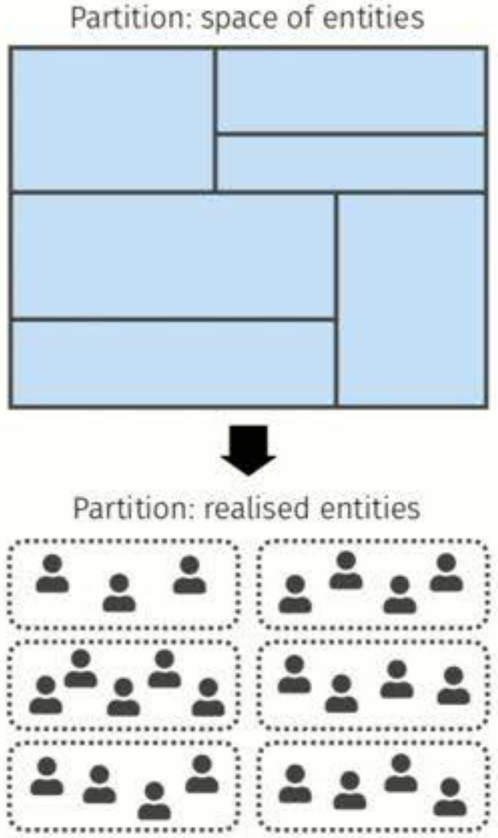
- Supports missing record values (MCAR)
- Supports arbitrary attribute similarity functions (one for each attribute)
- Incorporates a kind of "blocking": auxiliary partitions of the latent entities (not records)

Auxiliary partitions

- Enables distributed inference
- Crucially leaves the marginal posterior unchanged
- Partition the space of entities $\mathcal{V}_\otimes = \bigotimes_a \mathcal{V}_a$ using a deterministic function
  $\texttt{PartFn} : \mathcal{V}_\otimes \to \{1, \ldots, P\}$
- Update the model: now record r in table t is assigned to a partition γ_{tr}, then to an entity λ_{tr} within the partition

$$\gamma_{tr} | \mathbf{Y} \sim \text{Discrete}\left(\frac{|\mathcal{E}_1|}{|\mathcal{E}|}, \ldots, \frac{|\mathcal{E}_P|}{|\mathcal{E}|} \right)$$

$$\lambda_{tr} | \gamma_{tr} \sim \text{DiscreteUniform}(\mathcal{E}_{\gamma_{tr}})$$



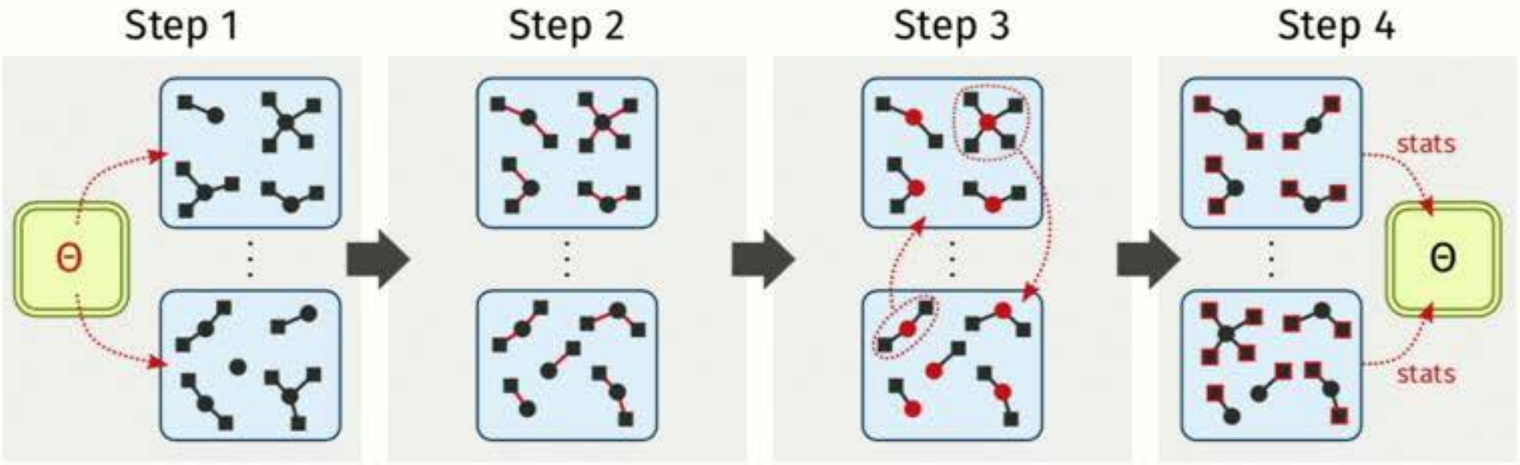
Distributed Markov chain Monte Carlo

Since the posterior for the linkage structure $p(\Lambda|X)$ is not tractable, we resort to approximate inference.

We propose an MCMC algorithm based on the partially-collapsed Gibbs framework (van Dyk and Park, 2008):

- regular Gibbs updates for the distortion probabilities θ_{ta}, distortion indicators z_{tra} and links λ_{tr}
- "marginalization" and "trimming" are applied to jointly update the entity attributes y_{ea} and the partition assignments for the linked records
- order of the updates is important (to preserve the stationary distribution)

Distributed Markov chain Monte Carlo



Update Θ on the master and broadcast to the workers.

Update Λ on the workers. Records may only link to entities within their assigned partitions.

Update Υ and Γ on the workers. Move the entities and records to their newly-assigned partitions.

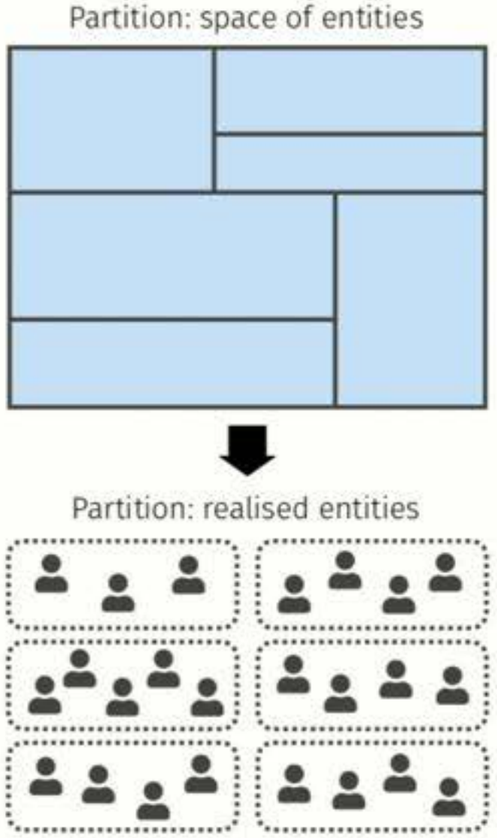
Update Z, then calculate summary stats on the workers. Broadcast to the master.

Auxiliary partitions

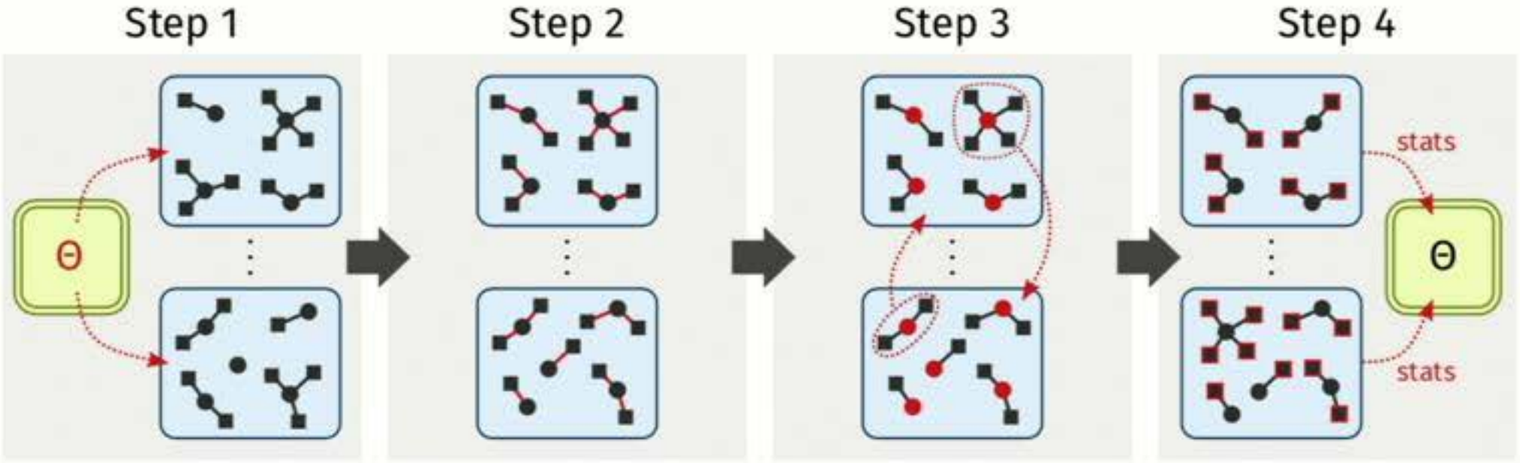
- Enables distributed inference
- Crucially leaves the marginal posterior unchanged
- Partition the space of entities $\mathcal{V}_{\otimes} = \bigotimes_a \mathcal{V}_a$ using a deterministic function
  $\texttt{PartFn} : \mathcal{V}_{\otimes} \to \{1, \ldots, P\}$
- Update the model: now record r in table t is assigned to a partition γ_{tr}, then to an entity λ_{tr} within the partition

$$\gamma_{tr}|\mathbf{Y} \sim \text{Discrete}\left(\frac{|\mathcal{E}_1|}{|\mathcal{E}|}, \ldots, \frac{|\mathcal{E}_P|}{|\mathcal{E}|}\right)$$

$$\lambda_{tr}|\gamma_{tr} \sim \text{DiscreteUniform}(\mathcal{E}_{\gamma_{tr}})$$



Distributed Markov chain Monte Carlo



Update Θ on the master and broadcast to the workers.

Update Λ on the workers. Records may only link to entities within their assigned partitions.

Update Υ and Γ on the workers. Move the entities and records to their newly-assigned partitions.

Update Ζ, then calculate summary stats on the workers. Broadcast to the master.

Tricks for speeding up inference

Two main bottlenecks:

1. linkage structure update $\mathcal{O}(\#\text{ records} \times \#\text{ entities})$
2. entity attribute update $\mathcal{O}(\#\text{ entities} \times \text{domain size})$

Tricks for speeding up inference

Two main bottlenecks:

1. linkage structure update $\mathcal{O}(\#\text{ records} \times \#\text{ entities})$
2. entity attribute update $\mathcal{O}(\#\text{ entities} \times \text{domain size})$

Solutions:

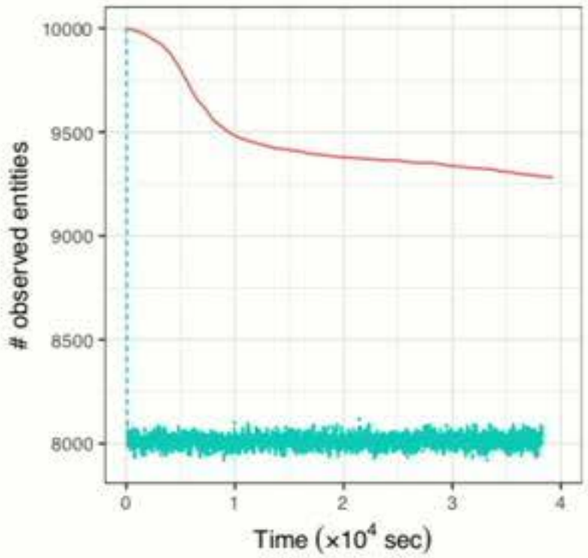
1. Indexing: Maintain indices from "entity attributes $\rightarrow$ entities" and "entities $\rightarrow$ linked records." This allows us to prune candidate links for a record
2. Thresholding similarity scores
3. Express the distribution for the entity attribute update as a two-component perturbation mixture model

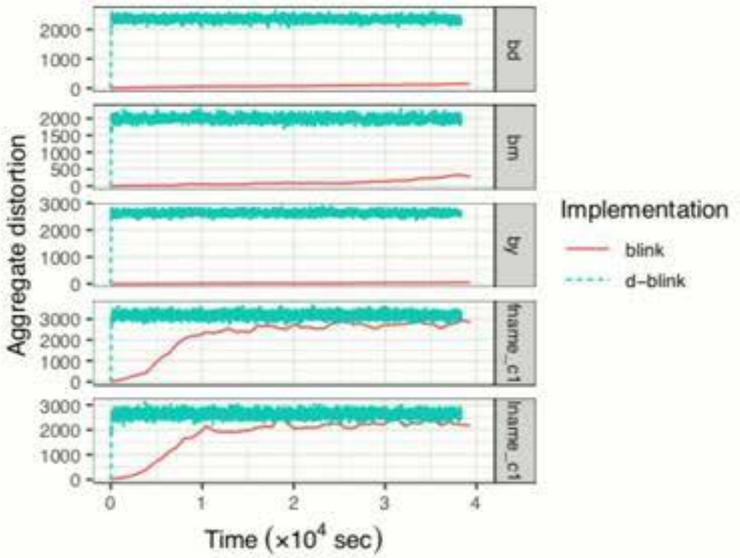
Experiments

- `ABSEmployee`. A synthetic data set used internally for linkage experiments by the ABS.
- `NCVR`. Two snapshots from the North Carolina Voter Registration database taken two months apart.
- `NLTCS`. A subset of the National Long-Term Care Survey comprising the 1982, 1989 and 1994 waves.
- `SHIW0810`. A subset from the Bank of Italy's Survey on Household Income and Wealth comprising the 2008 and 2010 waves.
- `RLdata10000`. A synthetic data set provided with the `RecordLinkage` R package.

Convergence of d-blink versus blink

We examined the rate of convergence of d-blink versus blink on RLdata10000 without partitioning.





d-blink converges rapidly, however blink fails to reach the equilibrium distribution within 11 hours.

Experiments

- Implemented `d-blink` and baselines in Apache Spark
- Ran experiments on a local server and Amazon EMR
- (Mostly) used a sample size of 10^3 after burnin (of 10^3 iterations) and thinning (keeping every 10th iteration)
- 3 real and 2 synthetic data sets

Data set	# records	# tables	# entities	# attributes	
				categorical	string
⋆ `ABSEmployee`	600,000	3	400,000	4	0
`NCVR`	448,134	2	296,433	3	3
`NLTCS`	57,077	3	34,945	6	0
`SHIW0810`	39,743	2	28,584	8	0
⋆ `RLdata10000`	10,000	1	9,000	2	3

Table: Assessment of the pairwise linkage performance for dblink and FS method as our baseline. We note that FS is supervised and does not propagate the entity resolution error exactly compared to dblink.[1]

Data set	Method	Pairwise measure		
		Precision	Recall	F1-score
`ABSEmployee`	dblink	**0.9943**	**0.8867**	**0.9374**
	Fellegi-Sunter (100)	0.9964	0.9510	0.9736
	Fellegi-Sunter (10)	0.4321	0.6034	0.9736
`NCVR`	dblink	**0.9179**	**0.9654**	**0.9411**
	Fellegi-Sunter (100)	0.8989	0.9974	0.9456
	Fellegi-Sunter (10)	0.8989	0.9974	0.9456
`NLTCS`	dblink	**0.8363**	**0.9102**	**0.8717**
	Fellegi-Sunter (100)	0.7969	0.9959	0.8853
	Fellegi-Sunter (10)	0.1902	0.9999	0.3196

[1] Comparisons to other semi-supervised methods are the same.

Posterior Bias Plot

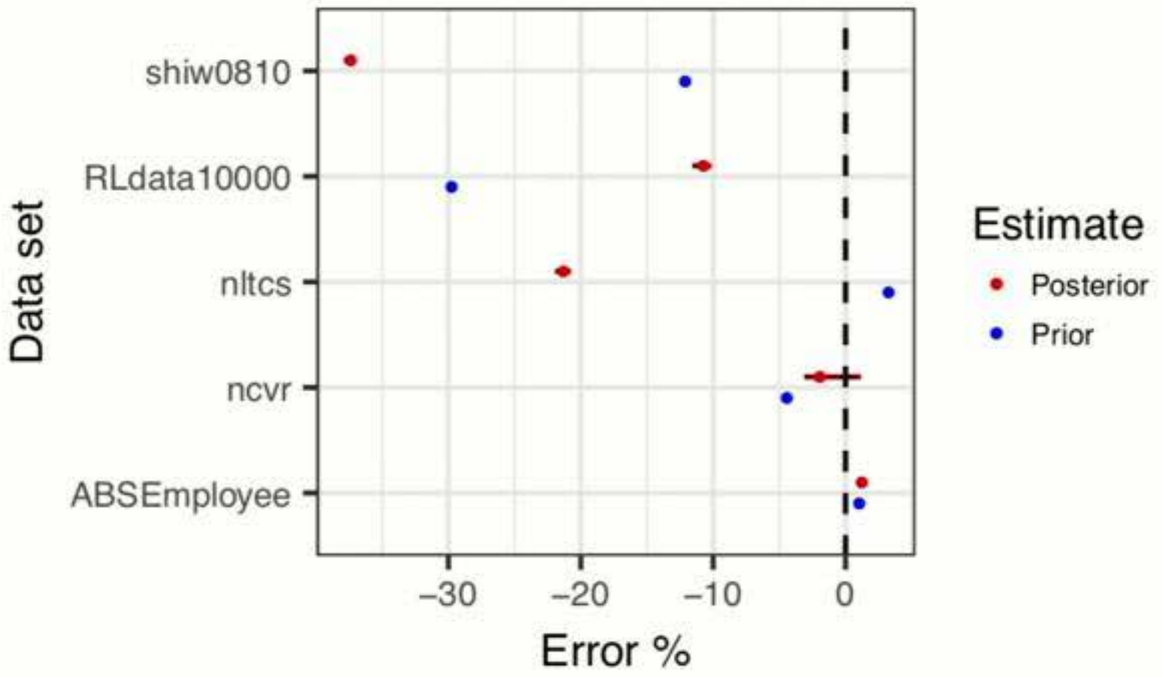
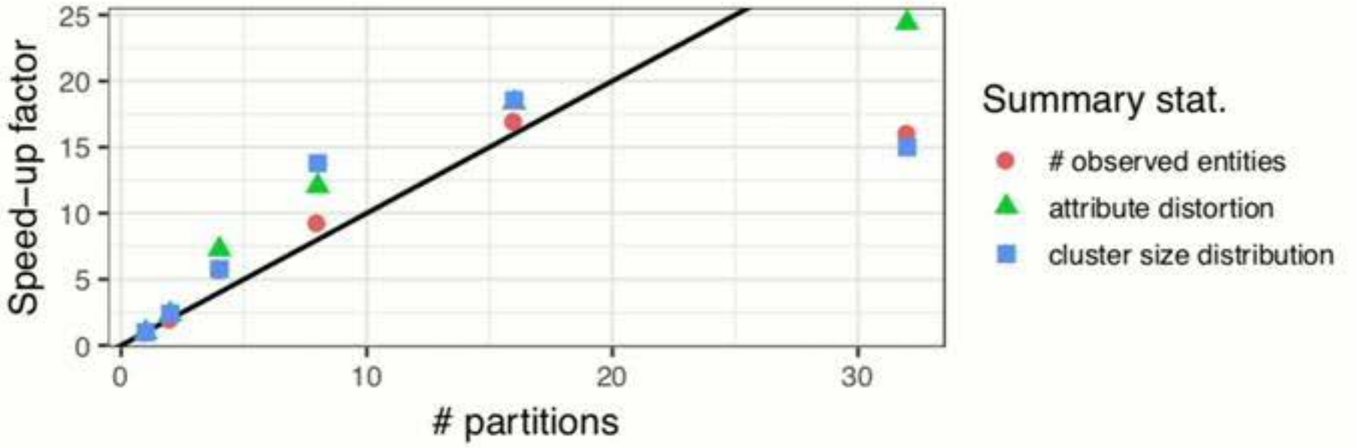


Figure: Error in the posterior and prior estimates for the number of observed entities for d-blink. The results show that the posterior estimate is very sharp and typically underestimates the true number, which is consistent with **RCS**, Hall, Fienberg (2016).

Does partitioning result in efficiency gains?

- Measure efficiency using ESS rate—the effective sample size generated per unit time
- Speed-up factor is the ESS rate relative to a baseline without partitioning
- Observe a near-linear speed-up for the `NLTCS` data set (tapering off beyond $\sim$ 20 partitions)



How does our approach handle data with little or no training data from a real application?

El Salvadoran Conflict

- El Salvador underwent a civil war from 1980 to 1991.
- The United Nations created a Truth Commission (UNTC) to record death casualties/disappearances related to the war by inviting witnesses through newspaper, radio, and television advertisements.
- Human rights groups often depend on accurate estimates and evaluations of the number of documented identifiable deaths for purposes of court ruling, etc.

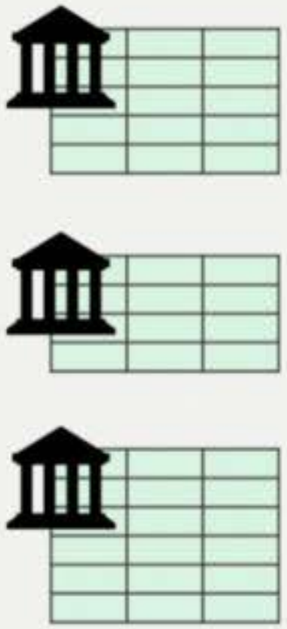
FUENTE DIRECTA

LISTA DE VICTIMAS CUYA IDENTIDAD NO SE MANTIENE EN RESERVA

APELLIDOS	NOMBRES	HECHO	FECHA	LUGAR	RESP1	RESP2	RESP3	RESP4
ABARCA PINEDA	ISABEL	DESAPARIC	0/ 6/81	80101	FFAA	FFAA		
ABARCA	JULIO CESAR	HOMICIDIO	10/ 7/84	60000				
ABARCA	LUIS	HOMICIDIO	14/ 5/80	42008	FFAA	PN	PARAMI	GN
ABARCA	LUIS	HOMICIDIO	20/ 1/82	100504	ESCUAD	FFAA		
ABARCA	MARIA CRUZ	VIOLACION	26/12/90	42101	FFAA			
ABARCA	MAURICIO	HOMICIDIO	0/ 3/88	60100	FFAA			
ABARCA	MILTON	HOMICIDIO	12/11/80	80118	PN	GN	FFAA	
ABARCA	NICOLAS ALFREDO	DESAPARIC	2/11/80	80100				
ABARCA	NICOLAS RUTILIO	HOMICIDIO	0/ 6/86	40000	FMLN			
ABARCA	RICARDO	HOMICIDIO	12/11/80	80118	PN	GN	FFAA	
ABARCA	ROSALINA	LESIONES	0/ 0/85	42802	FFAA			
ABARCA	RUFINO	HOMICIDIO	9/ 7/80	90605	PARAMI			
ABARCA ORELLANA	TOBIAS	HOMICIDIO	29/ 4/80	42102	PN	PARAMI		
ABARCA	TOVIAS	HOMICIDIO	29/ 4/80	42102				
ABARCA	ULALIO	HOMICIDIO	22/ 8/82	100502	FFAA	FFAA	FFAA	
ABELAR RONQUILLO	EDWIN ANTONIO	HOMICIDIO	13/ 1/86	20000	FFAA			
ABELAR	HERMINO	HOMICIDIO	13/ 1/82	60101	ESCUAD			
ABELAR	JOSE MARIO	HOMICIDIO	24/12/80	43300	FFAA			
ABREGO	ADRIAN	HOMICIDIO	16/ 5/80	40302	PARAMI	FFAA	GN	
ABREGO	ANDRES	HOMICIDIO	0/ 0/82	90205	GN			
ABREGO	ANTONIO	HOMICIDIO	10/ 8/83	40901	GN	PARAMI		
ABREGO	BENITO	HOMICIDIO	14/ 8/86	40200	FFAA			
ABREGO	BLANCA	HOMICIDIO	0/ 0/ 0	41401	PN			
ABREGO	CARLOS ALFREDO	DESAPARIC	29/11/80	16000				
ABREGO CASTRO	CARMEN	TORTURA	17/ 4/89	0	FFAA			
ABREGO	ELENA	HOMICIDIO	26/ 3/82	41902	FFAA			
ABREGO	FIDE	DESAPARIC	10/ 6/80	41501	GN	PARAMI		
ABREGO	FRANCISCO ANTONIO	HOMICIDIO	12/ 3/84	41902	PARAMI			
ABREGO	GUILLERMO	DESAPARIC	22/11/80	0				
ABREGO CASTRO	ISRAEL	HOMICIDIO	0/ 5/84	40906	GN	PARAMI		
ABREGO	JOSE	HOMICIDIO	24/ 2/85	71525	FFAA	FFAA		
ABREGO	JOSE ALFONSO	HOMICIDIO	11/11/80	40906	ESCUAD			
ABREGO CERAS	JOSE ERNESTO	DESAPARIC	22/11/80	0				
ABREGO CASTRO	JOSE MARINO DE JESUS	HOMICIDIO	2/11/89	60800	FFAA			
ABREGO NAVARRO		LESIONES	25/ 2/80	100107	FFAA	...	...	

Extract from Report of the UN Truth Commission of El Salvador (1993)

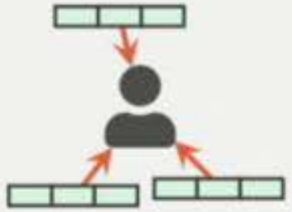
Data collected by multiple orgs



Data cleaning

normalisation and schema alignment

entity resolution



data fusion

Analysis leading to outputs

- statistics
- visualisations
- multiple systems estimation
- reports

Fully Non-parametric ER model

$$
\begin{aligned}
\sigma &\sim \text{Gamma}[b^0, b^1] \\
d &\sim \text{Beta}[c^0, c^1] \\
\boldsymbol{\pi} &\sim \text{PPY}[\sigma, d] \\
\lambda_{tr} \mid \boldsymbol{\pi} &\sim \text{Categorical}[\boldsymbol{\pi}] \\
\boldsymbol{y}_e | G_0 &\sim G_0 \\
H_{ea} | \mathrm{y}_{ea} &\sim \text{DP}[\beta_a; \Psi_a(\mathrm{y}_{ea})] \\
\theta_{ta} &\sim \text{Beta}[\alpha_a^0, \alpha_a^1] \\
\zeta_{tra} | \mathrm{y}_{\lambda_{tr} a}, \lambda_{tr} &= \max_{v \in V_a} e^{-\text{dist}_a(\mathrm{y}_{\lambda_{tr} a}, v)} \\
z_{tra} | \theta_{ta}, \zeta_{tra} &\sim \text{Bernoulli}[\theta_{ta} \zeta_{tra}] \\
x_{tra} | z_{tra}, y_{\lambda_{tr} a}, \lambda_{tr} &\sim (1 - z_{tra}) \delta[y_{\lambda_{tr} a}] + z_{tra} H_{ea}
\end{aligned}
$$

where $g_0\ \boldsymbol{y} = \prod_a \phi_a(v)$ and $\psi_a(v|\mathrm{y}) \propto \phi_a(v) e^{-\text{dist}_a(y,v)}$.

FUENTE DIRECTA

LISTA DE VICTIMAS CUYA IDENTIDAD NO SE MANTIENE EN RESERVA

APELLIDOS	NOMBRES	HECHO	FECHA	LUGAR	RESP1	RESP2	RESP3	RESP4
ABARCA PINEDA	ISABEL	DESAPARIC	0/ 6/81	80101	FFAA	FFAA		
ABARCA	JULIO CESAR	HOMICIDIO	10/ 7/84	60000				
ABARCA	LUIS	HOMICIDIO	14/ 5/80	42008	FFAA	PN	PARAMI	GN
ABARCA	LUIS	HOMICIDIO	20/ 1/82	100504	ESCUAD	FFAA		
ABARCA	MARIA CRUZ	VIOLACION	26/12/90	42101	FFAA			
ABARCA	MAURICIO	HOMICIDIO	0/ 3/88	60100	FFAA			
ABARCA	MILTON	HOMICIDIO	12/11/80	80118	PN	GN	FFAA	
ABARCA	NICOLAS ALFREDO	DESAPARIC	2/11/80	80100				
ABARCA	NICOLAS RUTILIO	HOMICIDIO	0/ 6/86	40000	FMLN			
ABARCA	RICARDO	HOMICIDIO	12/11/80	80118	PN	GN	FFAA	
ABARCA	ROSALINA	LESIONES	0/ 0/85	42802	FFAA			
ABARCA ORELLANA	RUFINO	HOMICIDIO	9/ 7/80	90605	PARAMI			
ABARCA	TOBIAS	HOMICIDIO	29/ 4/80	42102	PN	PARAMI		
ABARCA	TOVIAS	HOMICIDIO	29/ 4/80	42102				
ABARCA	ULALIO	HOMICIDIO	22/ 8/82	100502	FFAA	FFAA	FFAA	
ABELAR RONQUILLO	EDWIN ANTONIO	HOMICIDIO	13/ 1/86	20000	FFAA			
ABELAR	HERMINO	HOMICIDIO	13/ 1/82	60101	ESCUAD			
ABELAR	JOSE MARIO	HOMICIDIO	24/12/80	43300	FFAA			
ABREGO	ADRIAN	HOMICIDIO	16/ 5/80	40302	PARAMI	FFAA	GN	
ABREGO	ANDRES	HOMICIDIO	0/ 0/82	90205	GN			
ABREGO	ANTONIO	HOMICIDIO	10/ 8/83	40901	GN	PARAMI		
ABREGO	BENITO	HOMICIDIO	14/ 8/86	40200	FFAA			
ABREGO	BLANCA	HOMICIDIO	0/ 0/ 0	41401	PN			
ABREGO CASTRO	CARLOS ALFREDO	DESAPARIC	29/11/80	16000				
ABREGO	CARMEN	TORTURA	17/ 4/89	0	FFAA			
ABREGO	ELENA	HOMICIDIO	26/ 3/82	41902	FFAA			
ABREGO	FIDE	DESAPARIC	10/ 6/80	41501	GN	PARAMI		
ABREGO	FRANCISCO ANTONIO	HOMICIDIO	12/ 3/84	41902	PARAMI			
ABREGO CASTRO	GUILLERMO	DESAPARIC	22/11/80	0				
ABREGO	ISRAEL	HOMICIDIO	0/ 5/84	40906	GN	PARAMI		
ABREGO	JOSE	HOMICIDIO	24/ 2/85	71525	FFAA	FFAA		
ABREGO CERAS	JOSE ALFONSO	HOMICIDIO	11/11/80	40906	ESCUAD			
ABREGO CASTRO	JOSE ERNESTO	DESAPARIC	22/11/80	0				
ABREGO NAVARRO	JOSE MARINO DE JESUS	HOMICIDIO	2/11/89	60800	FFAA			
		LESIONES	25/ 2/80	100107	FFAA			

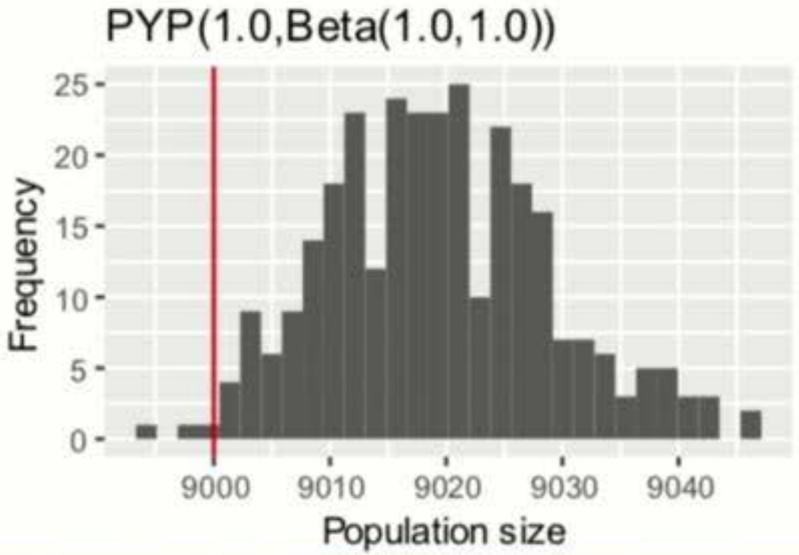
Extract from Report of the UN Truth Commission of El Salvador (1993)

Fully Non-parametric ER model

$$
\begin{aligned}
\sigma &\sim \text{Gamma}[b^0, b^1] \\
d &\sim \text{Beta}[c^0, c^1] \\
\boldsymbol{\pi} &\sim \text{PPY}[\sigma, d] \\
\lambda_{tr} \mid \boldsymbol{\pi} &\sim \text{Categorical}[\boldsymbol{\pi}] \\
\boldsymbol{y}_e | G_0 &\sim G_0 \\
H_{ea} | \mathrm{y}_{ea} &\sim \text{DP}[\beta_a; \Psi_a(\mathrm{y}_{ea})] \\
\theta_{ta} &\sim \text{Beta}[\alpha_a^0, \alpha_a^1] \\
\zeta_{tra} | \mathrm{y}_{\lambda_{tr} a}, \lambda_{tr} &= \max_{v \in V_a} e^{-\text{dist}_a(\mathrm{y}_{\lambda_{tr} a}, v)} \\
z_{tra} | \theta_{ta}, \zeta_{tra} &\sim \text{Bernoulli}[\theta_{ta} \zeta_{tra}] \\
x_{tra} | z_{tra}, y_{\lambda_{tr} a}, \lambda_{tr} &\sim (1 - z_{tra}) \delta[y_{\lambda_{tr} a}] + z_{tra} H_{ea}
\end{aligned}
$$

where $g_0\ \boldsymbol{y} = \prod_a \phi_a(v)$ and $\psi_a(v|\mathrm{y}) \propto \phi_a(v) e^{-\text{dist}_a(y,v)}$.

RLdata10000



Prior type	% rel. error # entities	Precision	Recall
PYP	0.2%	0.97	0.98
DP	-50%	0.07	1.00
Uniform	-47%	0.10	0.99
Coupon	-15%	0.50	0.99

Results

Table: Comparison of PYP, DP and Uniform prior on UNTC.

Prior	a	b	ϑ	σ	Precision	Recall	Posterior mean	SE	Runtime (sec)
PYP	1	99	1.7272	0.9890	**0.900**	0.153	725.45	1.27	892.17
			2.5663	0.9885	0.900	0.153	725.58	1.64	894.91
			4.6017	0.9875	0.900	0.153	728.21	1.27	803.36
DP	1	99	1	-	0.770	0.797	678.237	1.38	997.8
			2	-	**0.797**	**0.797**	680.08	1.79	1054.2
			3	-	0.793	0.780	682.18	1.74	1042.2
Uniform	1	73.5	-	-	**0.867**	0.661	692.47	2.58	3490.09
	1	99	-	-	0.826	0.644	688.84	2.18	3280.19

Example of False Positive Error

Table: Our model currently clusters these records to the same latent entity. But according to the hand matched labels, there are actually two latent entities here. We believe that this hand-matched label could be easily labeled as match or not-match depending on the hand-matcher.

Firstname	Lastname	Year	Month	Day	Department	Municipality
CARMEN	ALFARO	1982	3	21	7	716
JOSE	ALFARO GAMES	1980	3	22	7	716
CARMEN	ALFARO GAMES	1980	3	22	7	716

UNTC dataset

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

We utilize a similar type of framework to Marchant et. al (2019), however we consider a set of more flexible priors on the linkage structure and a different type of string metrics for the hispanic names.

Theoretical Limits for Entity Resolution

Are there theoretical limits for entity resolution? Yes!

Two very different approaches:

1. Johndrow, Lum, and Dunson, Biometrika, 2018.
   - The authors show that that for p features or when the number of entities is small compared to the number of records that entity resolution is "effectively impossible."
2. **RCS**, Barnes, Neiswanger (2017), AIStats, `http://proceedings.mlr.press/v54/`.
   - We derive performance bounds (lower bounds) using the KL divergence on when a latent entity is mis-classified, showing when the bounds are tight.

Thank you!

Questions?

Contact: beka@stat.duke.edu

`https://github.com/resteorts/record-linkage-tutorial`