Entity resolution with an application to the El Salvadoran conflict

Rebecca C. Steorts
Assistant Professor, Duke University
Department of Statistical Science, Computer Science, Biostatistics and Bioinformatics, the information initiative at Duke,
and the Social Science Research Institute

Joint work with Bihan Zhuang (Duke), Neil Marchant, and Ben Rubinstein (Melbourne)

Human rights conflict

The New York Times

SCIENCE

New Estimate Raises Civil War Death Toll

By GUY GUGLIOTTA APRIL 2, 2012

REUTERS

ENERGY

APRIL 8, 2019 / 4:57 AM / 8 DAYS AGO

Libyan death toll rises as battle for Tripoli intensifies

Ahmed Elumami, Ayman al-Warfalli

The Washington Post
Democracy Dies in Darkness

WorldViews

The Syrian war's death toll is absolutely staggering. But no one can agree on the number.

By **Adam Taylor**
March 15, 2016

AP

AP NEWS

Top Stories Topics Video Listen

Nearly 400,000 'excess deaths' caused by South Sudan war

By SAM MEDNICK September 26, 2018

CNN World

UN verifies deaths and injuries of 7,000 children in Syrian war

By Angela Dewan, CNN

Posted at 5:35 AM ET, Sat July 28, 2018

Challenges in human rights

- Data is becoming rapidly available at our fingertips

- Data quality issues must be managed
  - Typos
  - Transcription errors
  - Missing data
  - Multiple data sources
  - Convenient samples

Case study: El Salvador

- El Salvador underwent a civil war from 1980 to 1991.
- The United Nations created a Truth Commission (UNTC) to record death casualties and 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.

UNTC data set

- Due to the data collection process, the UNTC data set is subject to noise, distortions, missing values, and duplication.
- Other challenges of this data set include:
    - lack of unique identifiers
    - hard to construct training data set
    - complexity of Spanish names

UNTC data set

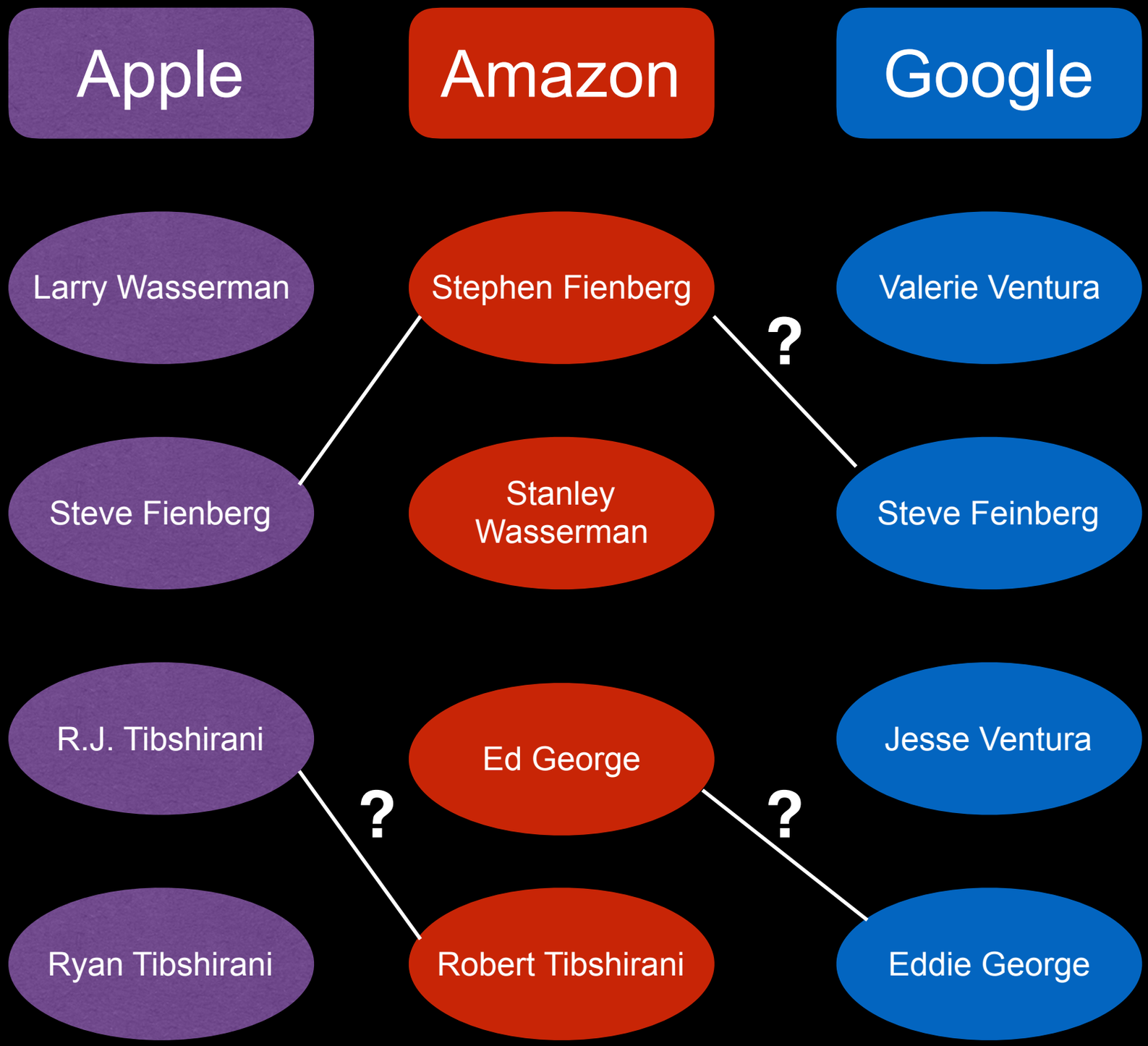
Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Figure 1: The first the records most likely refer to the same person, name JOSE FLORES. However, it's unclear if the fourth and fifth records refer to the same person or two different people. Is this the same person that died on the same day? Or are these a father and son that died in the same household?

Entity resolution

Entity resolution (record linkage or de-duplication) is the process of merging together noisy databases to remove duplicate entities, often in the absence of a unique identifier.

Fellegi and Sunter (1969) Christen (2012)


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Valerie Ventura
?
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R.J. Tibshirani
Ed George
Jesse Ventura
?
?
Ryan Tibshirani
Robert Tibshirani
Eddie George

Larry Wasserman

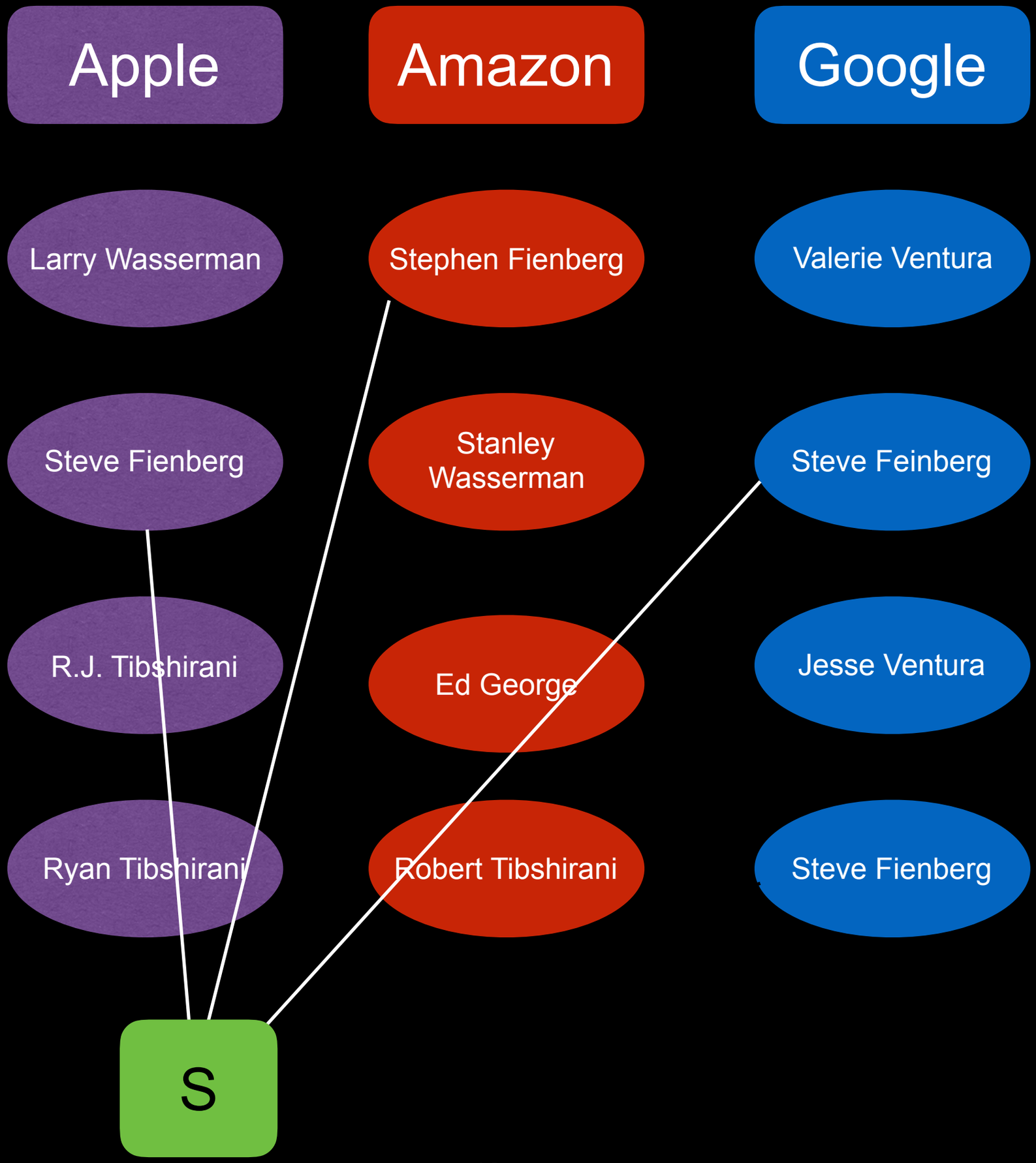
Larry Wasserman
1014 Murray Hill Avenue
Pittsburgh, PA 15217
412-361-3146

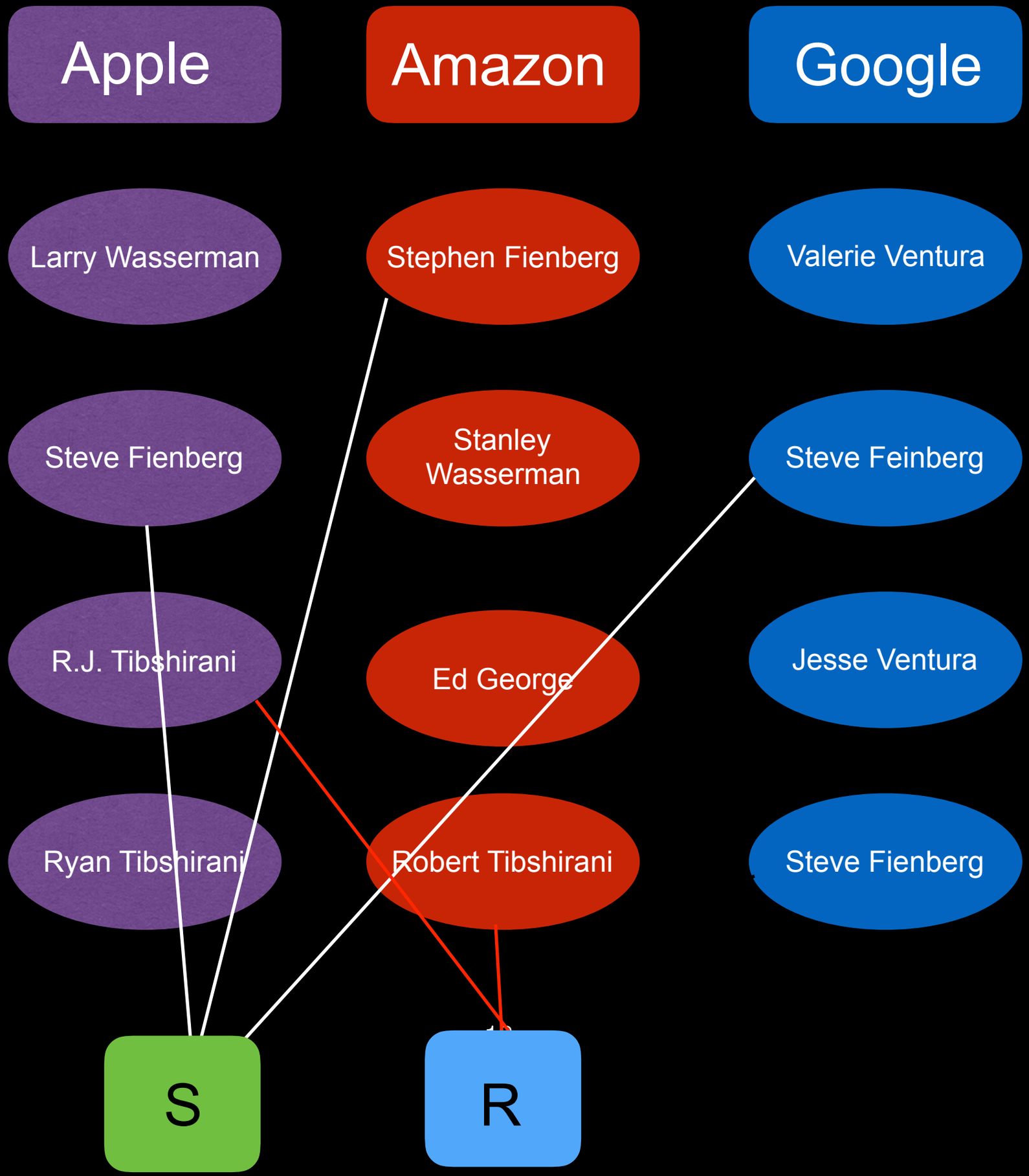
Steve Feinberg

240 Collins Drive
Pittsburgh, PA
50-54
412-793-3313

Stephen Fienberg

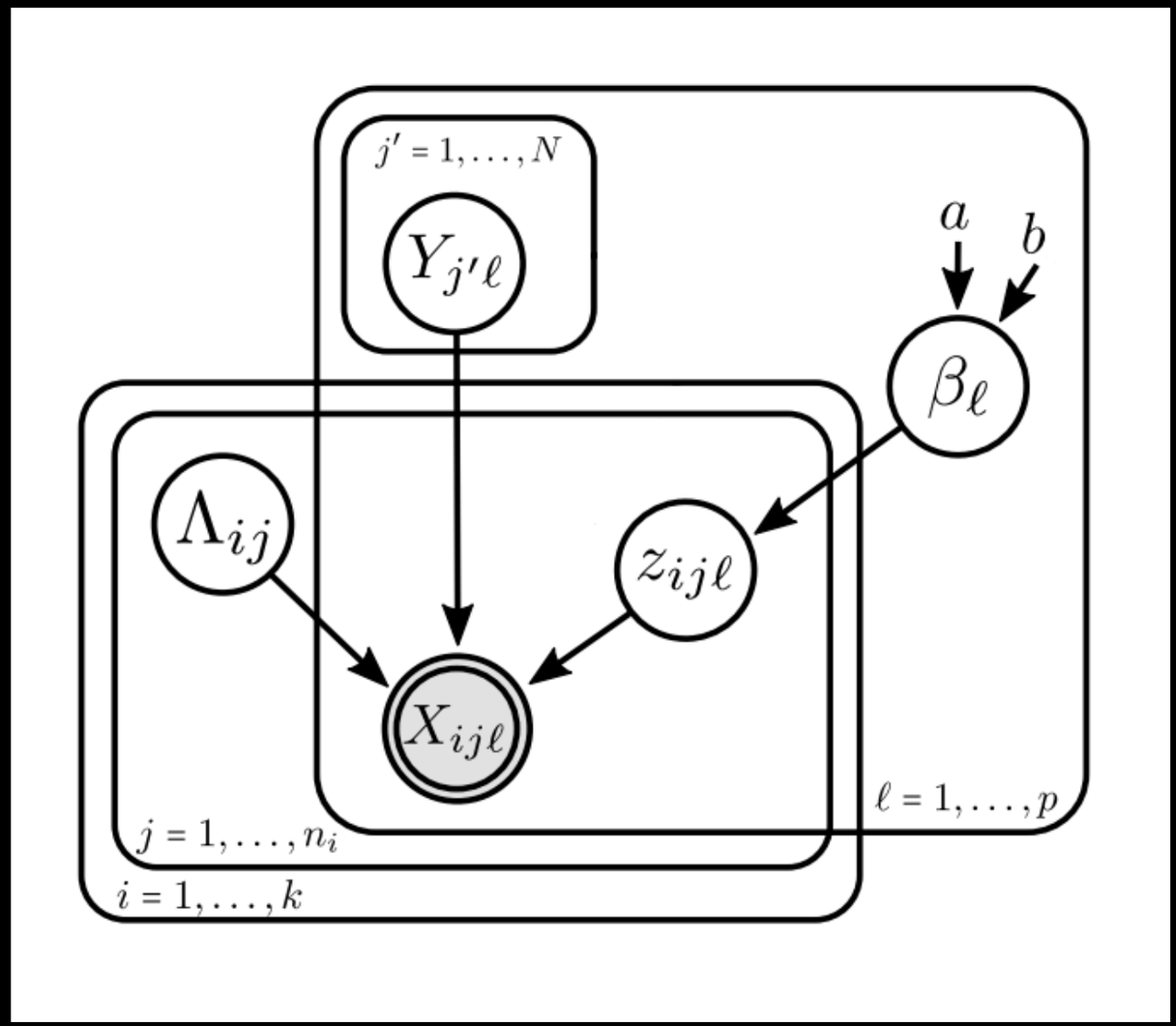
537 N Neville Street
Pittsburgh, PA 15213
65+
412-683-5599


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Valerie Ventura
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R.J. Tibshirani
Ed George
Jesse Ventura
Ryan Tibshirani
Robert Tibshirani
Steve Fienberg
S


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Valerie Ventura
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R.J. Tibshirani
Ed George
Jesse Ventura
Ryan Tibshirani
Robert Tibshirani
Steve Fienberg
S
R

Background

- Our work builds off Copas and Hilton and Tancredi and Liseo (2011).
- It then builds off Steorts, Hall Fienberg (2016) and Steorts (2015), and Sadinle (2014).



Why this approach?

- It supports multiple databases (tables or files).
- It supports both categorical and text-based data structures.
- The distortion in attributes can be informed by similarity functions.
- It models the entities as latent variables, which ensures transitivity and provides a merged representation of linked records.
- Uncertainty of the entity resolution process can be quantified exactly.
- Our approach is scalable using partially collapsed Gibbs and truncated similarity metrics.
- This approach has provable performance bounds under the Kullback–Leibler divergence.

Steorts (2015); Steorts, Barnes, and Neiswanger (2017), Marchant, Steorts, Kaplan, Rubenstein, Elazar (2019)

Our contribution

- We propose Bayesian nonparametric priors on the linkage structure. Specifically, we consider the Pitman Yor Process (PYP) prior and Dirichlet Process (DP) prior.

- We incorporate missing data scheme into the proposed model.

- We do not require any dimensionality reduction (such as blocking), which means that the only sources of error in our inferential method comes from the data and the entity resolution task. (This differs from the original work of Sadinle (2014) where blocking is required for computationally scalability).

- We derive the full conditional distributions for this alternative model and implement the Gibbs sampler.

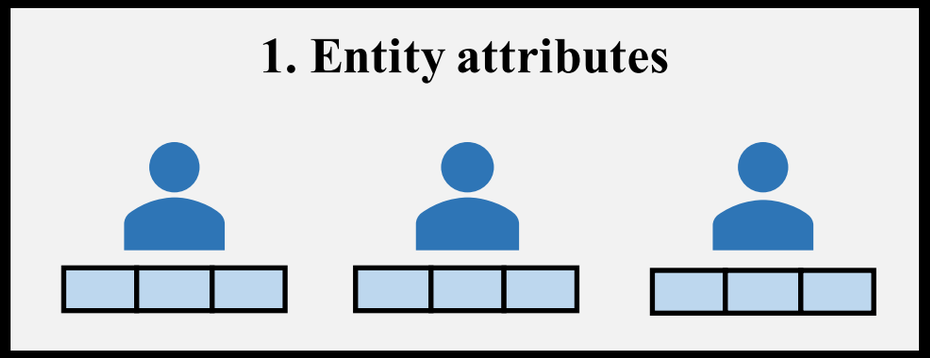
- We provide speeds up as Gibbs sampling is quite slow using partially collapsed Gibbs and truncation metrics.

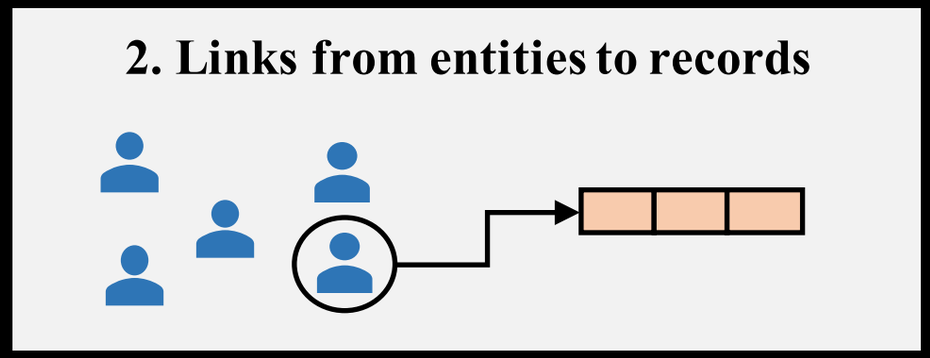
- We propose an asymmetric similarity metric for the UNTC data set.

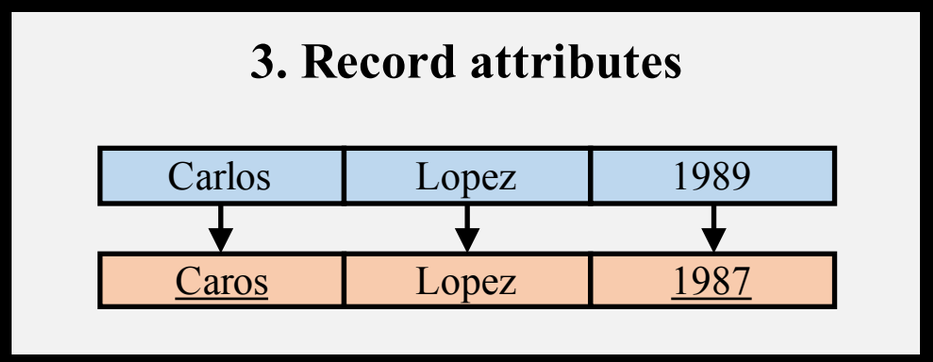
- We provide evaluations on both synthetic and the UNTC data.

Bayesian model for entity resolution

- We model the generative process of the records.
- There are three components to model.

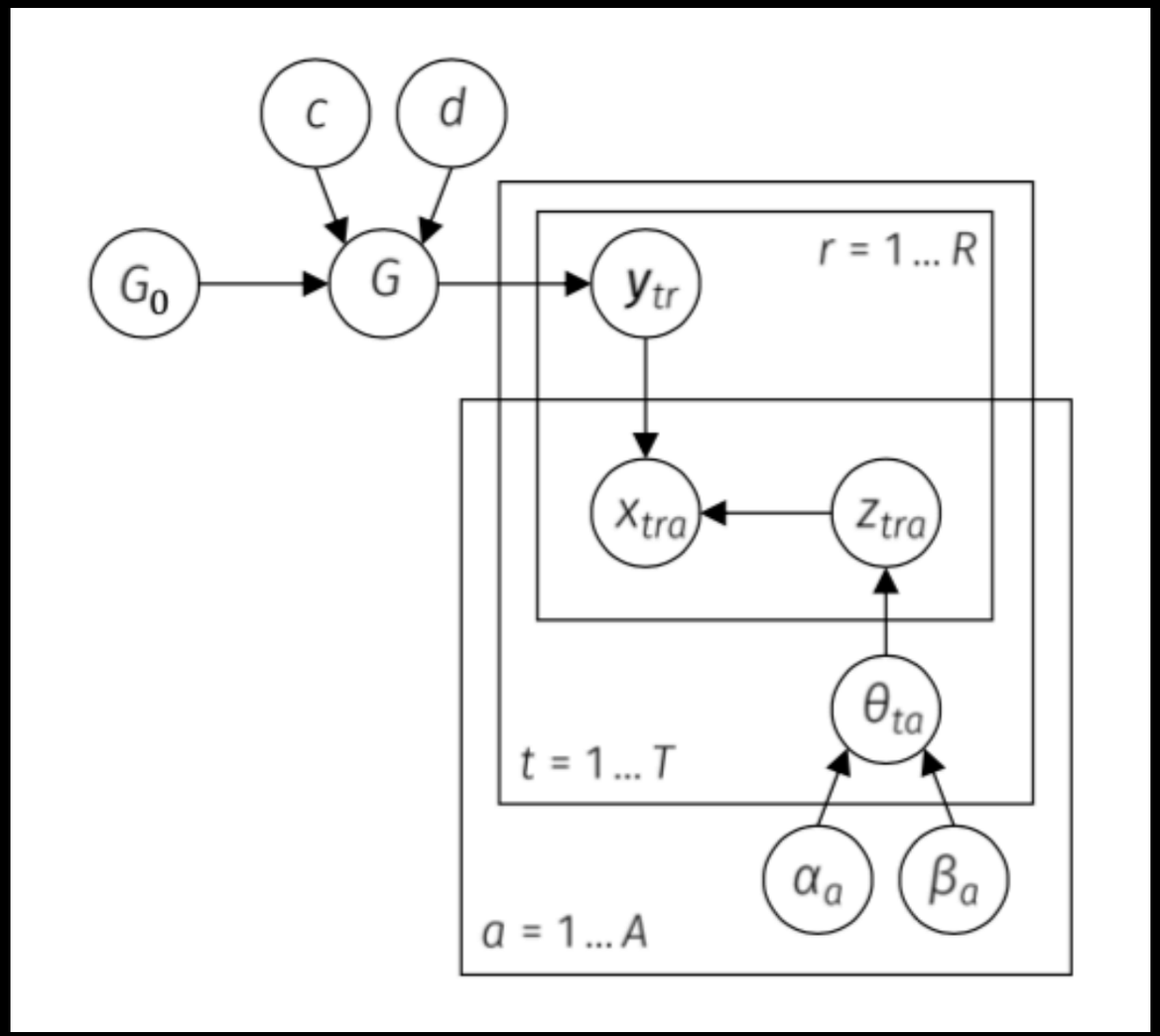






Bayesian model for entity resolution

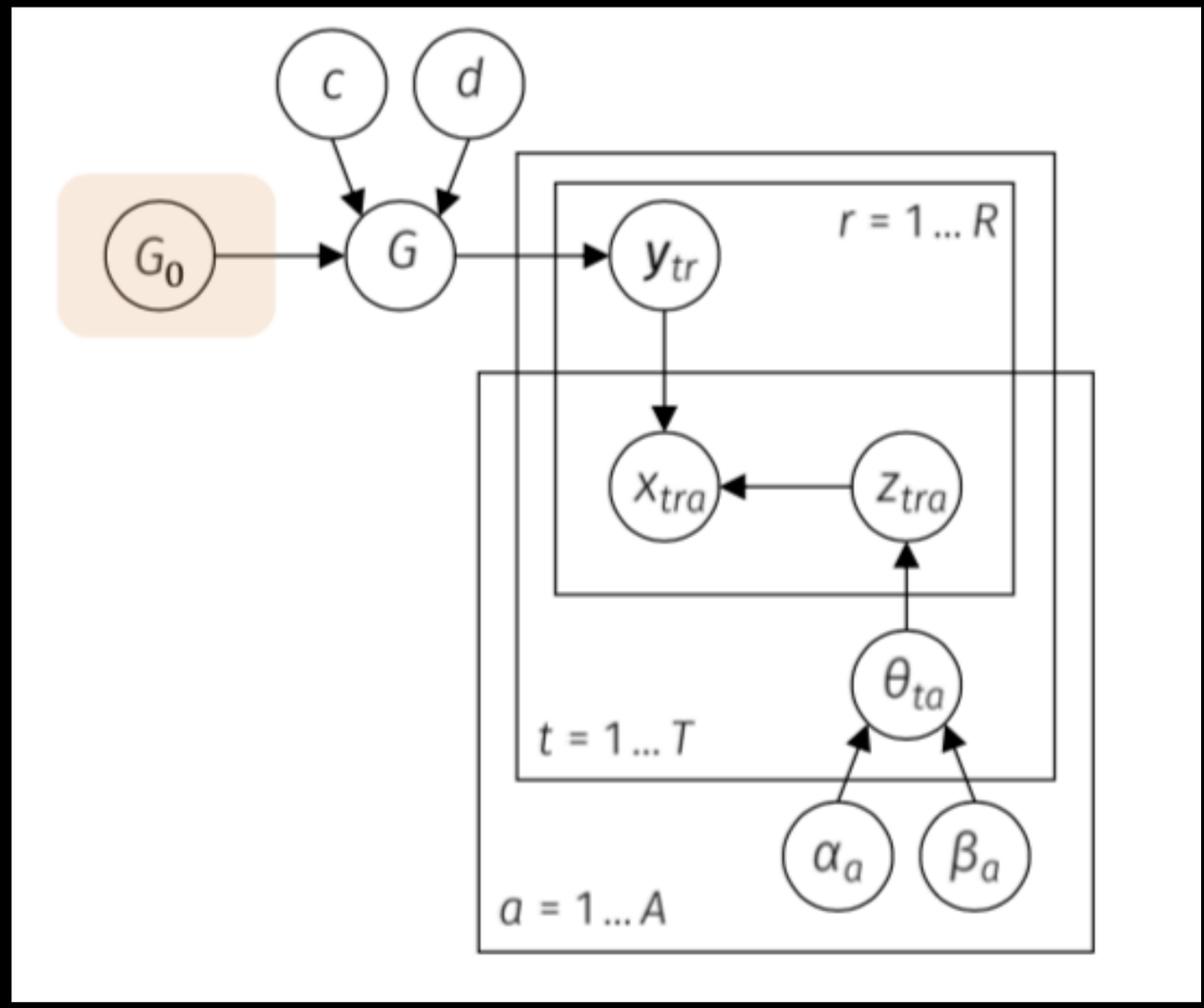
1. Entity attributes
   - Use the empirical distribution function
   - Assume attributes are independent
2. Links from entities to attributes
   - Consider BNP priors
3. Record attributes
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



a: attribute, r: record, t: table

Bayesian model for entity resolution

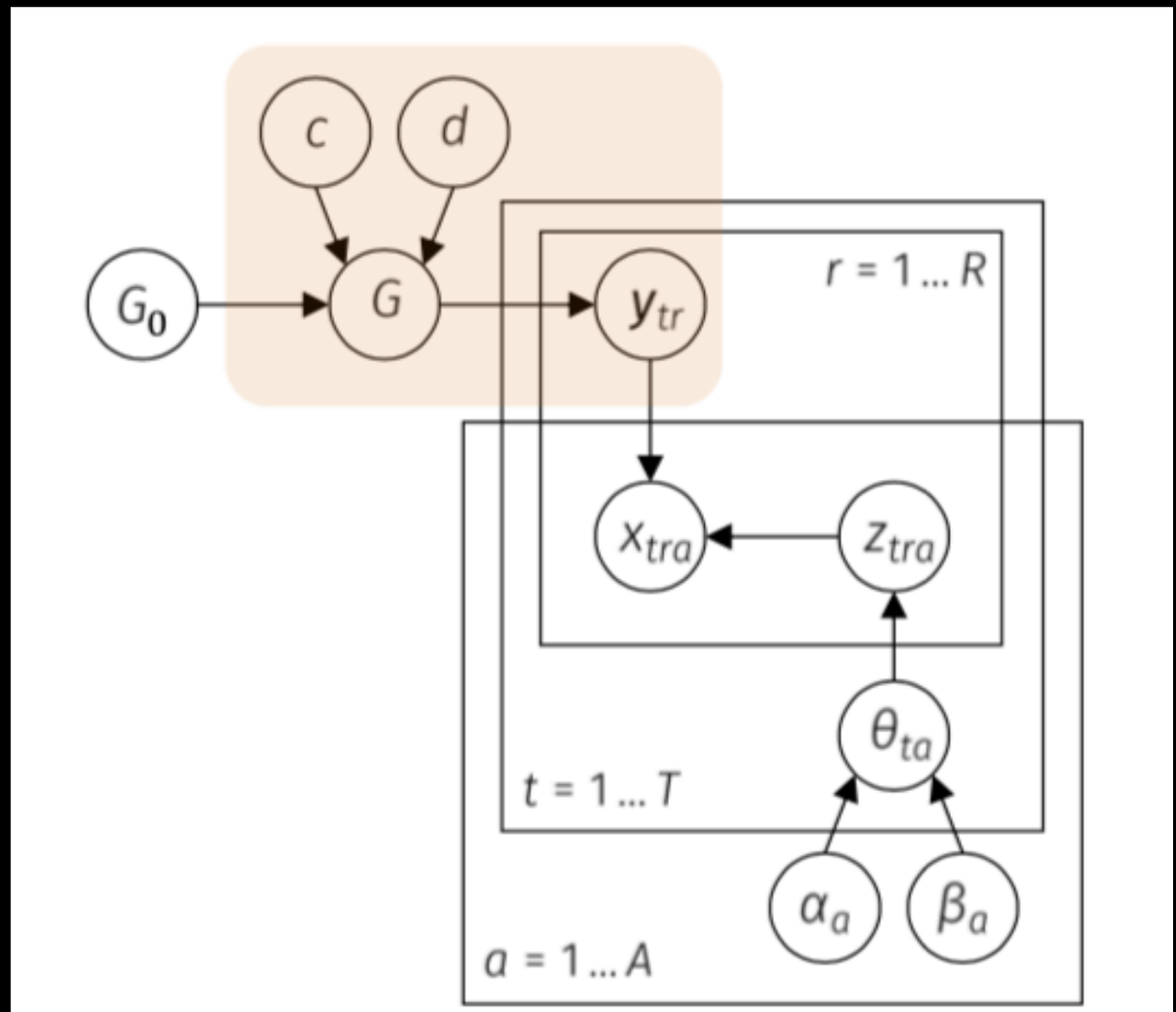
1. **Entity attributes**
   - Use the empirical distribution function
   - Assume attributes are independent
2. **Links from entities to attributes**
   - Consider BNP priors
3. **Record attributes**
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



a: attribute, r: record, t: table

Bayesian model for entity resolution

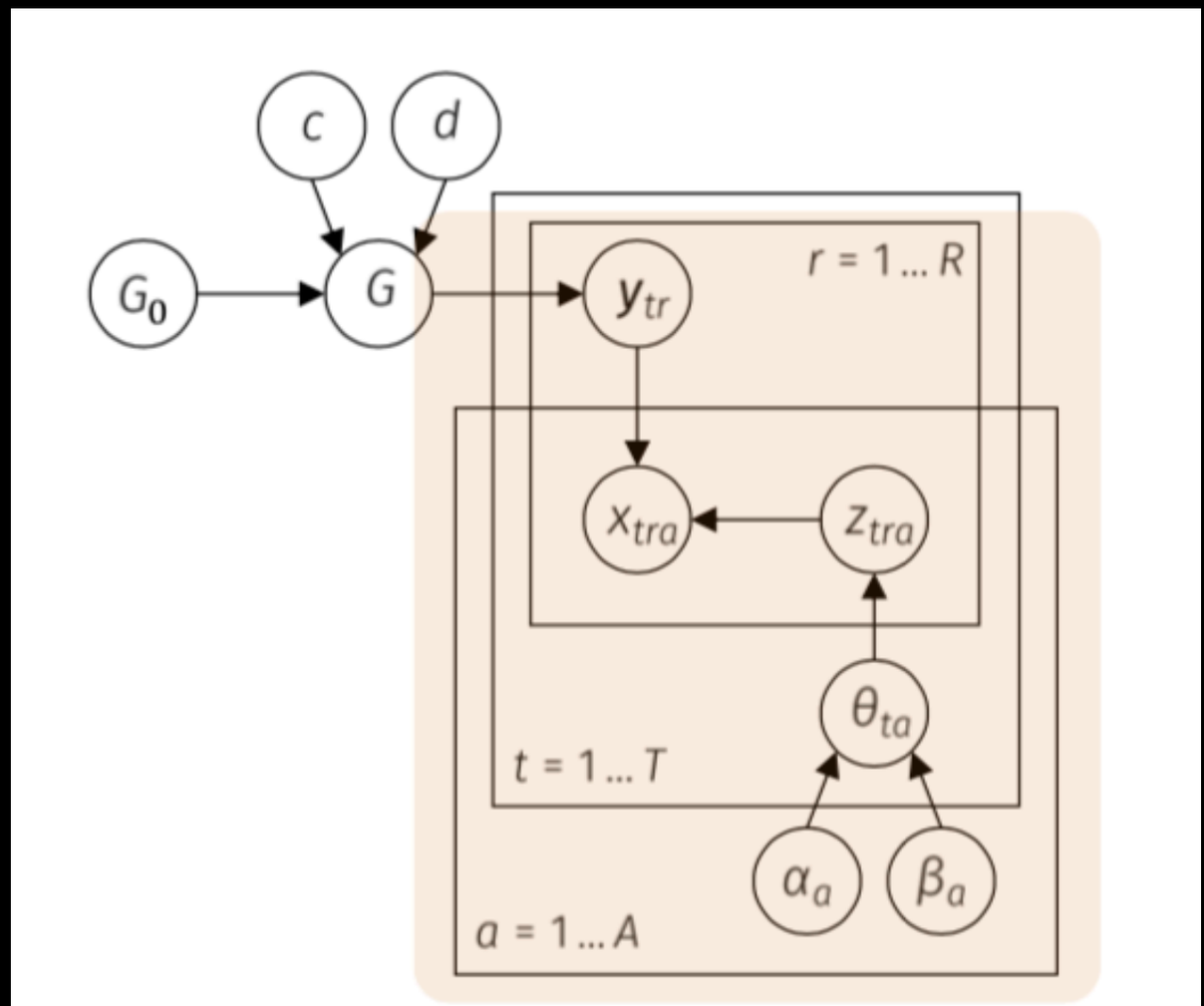
1. **Entity attributes**
   - Use the empirical distribution function
   - Assume attributes are independent
2. **Links from entities to attributes**
   - Consider BNP priors
3. **Record attributes**
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



a: attribute, r: record, t: table

Bayesian model for entity resolution

1. **Entity attributes**
   - Use the empirical distribution function
   - Assume attributes are independent
2. **Links from entities to attributes**
   - Consider BNP priors
3. **Record attributes**
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value

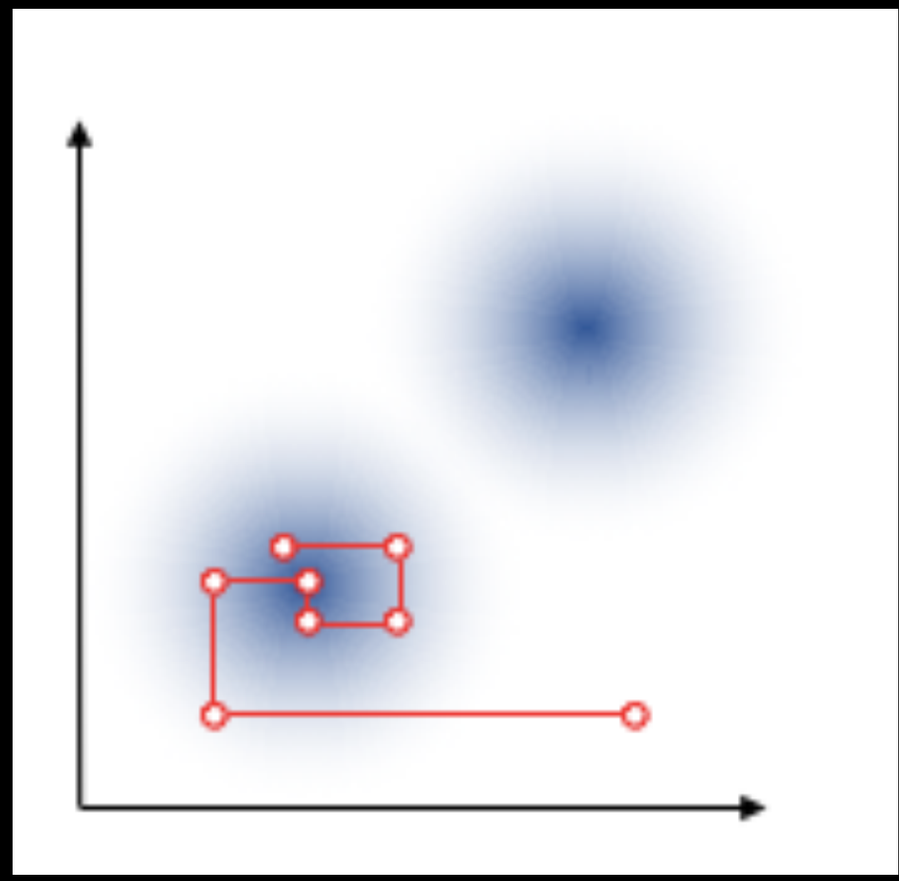


a: attribute, r: record, t: table

Gibbs sampler

Reduce the problem to a sequence of low-dimensional simulations:

- Usually update one variable at a time, holding all others fixed
- Conditional distributions must be known and easy to sample from



Details for this model:

- Partially-collapse the distortion indicators to improve mixing
- Need to introduce auxiliary variables to update the hyper-priors

Gibbs sampling tricks

Naïve approach scales poorly:

- Linkage structure update
  $O(\#\text{ entities} \times \#\text{ records})$
- Entity attribute update
  $O(\#\text{ entities} \times \text{domain size})$

1. Indexing

- Maintain indexes from entity attributes → entities; entities → linked records
- Prepare candidate links using multiple set intersection

2. Perturbation sampling

- Write entity attribute distribution as a two-component mixture
- Perturbation component has a large weights and much smaller support

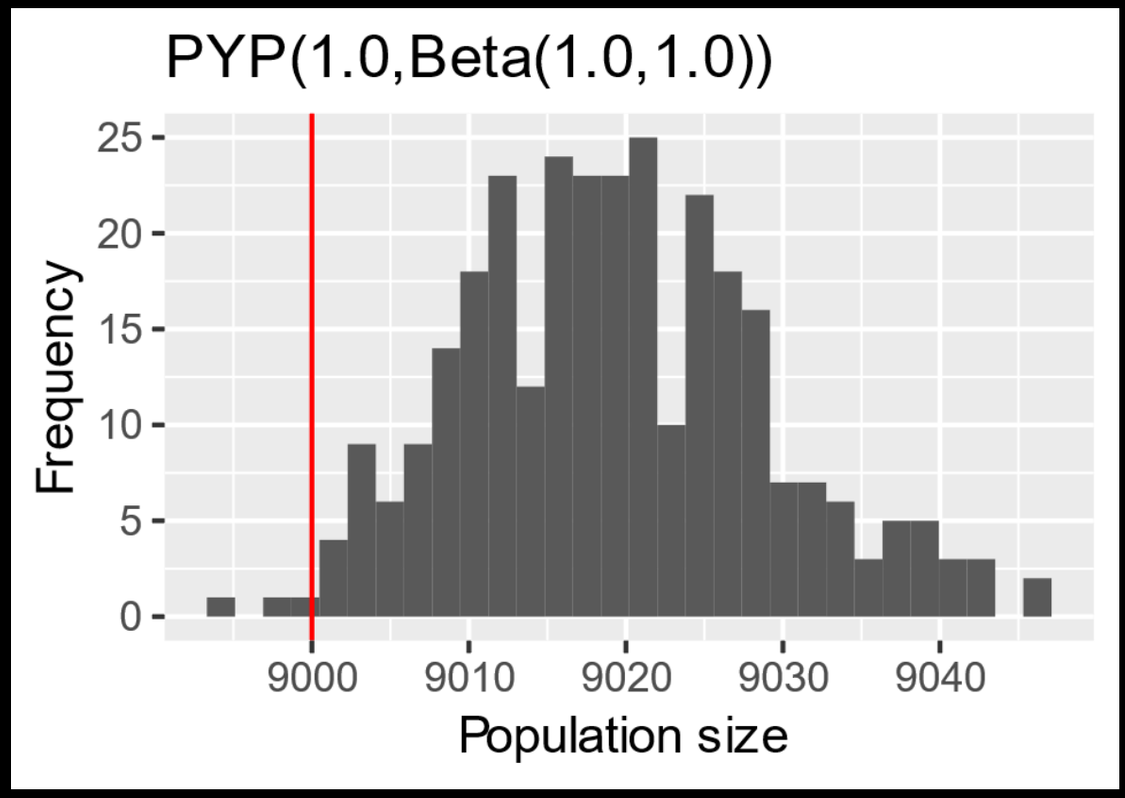
3. Similarity threshold

- Similarities that are below a given threshold as assumed to be completely dissimilar
- Higher threshold means more efficient

Experiment: Synthetic data

Synthetic “RLdata10000”:
- 10k records, 9k true entities
- Attributes: first name, last name, DOB
- Ground truth available Parameter settings

Parameter settings:
- Normalized Levenshtein similarity for names
- Constant similarity for DOB
- Weak prior for low distortion Beta(1,100)
- 500 burn-in + 3000 iterations



prior	% rel. error	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

El Salvadoran Case Study

El Salvador is divided into 14 departments for administrative purposes, subdivided into 262 municipalities.

El Salvadoran Case Study

List of victims from UNTC report
- 5395 scanned records
- No definitive ground truth
- Rough hand labels for two departments:
  Cuscatlán & Ahuachapán (Sadinle, 2014)

Parameter settings
- Names: custom string similarity
- DOD: similarity based on numeric absolute difference
- Muni/dept: similarity based on geographic proximity
- 10k burn-in + 40k iterations

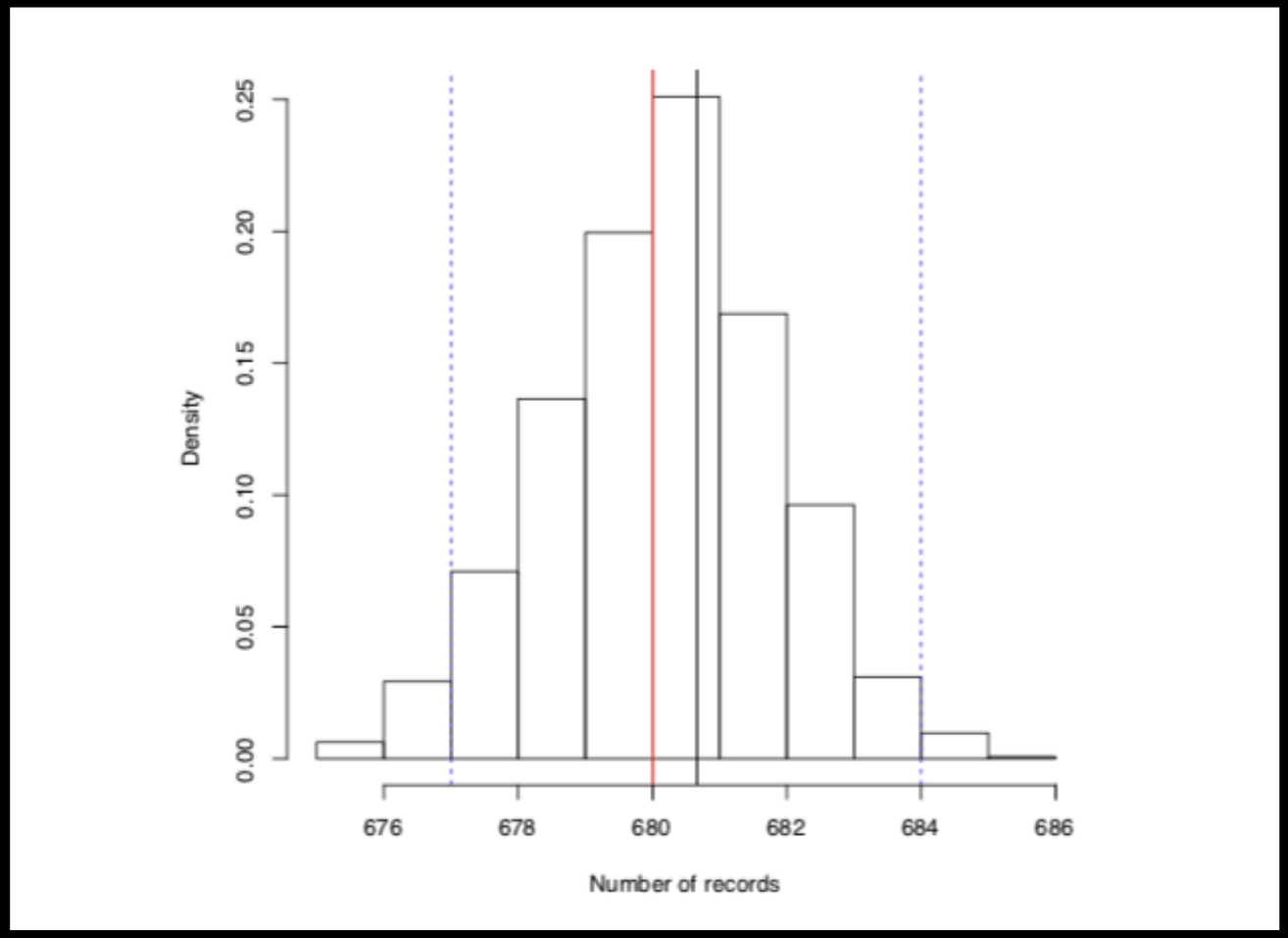
FUENTE DIRECTA
LISTA DE VICTIMAS CUYA IDENTIDAD NO SE MANTIENE EN RESERVA

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	
ABARCA	EDWIN ANTONIO	HOMICIDIO	13/ 1/86	20000	FFAA			
ABELAR RONQUILLO	HERMINO	HOMICIDIO	13/ 1/82	60101	ESCUAD			
ABELAR	JOSE MARIO	HOMICIDIO	24/12/80	43300	FFAA			
ABELAR	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	JOSE ALFONSO	HOMICIDIO	11/11/80	40906	ESCUAD			
ABREGO DERAS	JOSE ERNESTO	DESAPARIC	22/11/80	0				
ABREGO CASTRO	JOSE MARINO DE JESUS	HOMICIDIO	2/11/89	60800	FFAA			
ABREGO NAVARRO	JOSE VICTOR	LESIONES	25/ 2/80	100107	FFAA			
ABREGO	LEONICIO	HOMICIDIO	22/10/82	71524	ESCUAD	PN	PARAMI	
ABREGO	LINORA	HOMICIDIO	4/ 8/82	100107	FFAA			
ABREGO URILLAS	LUIS	HOMICIDIO	0/ 0/ 0	40900	FFAA	FFAA		
ABREGO	LUIS FELIPE	DESAPARIC	0/11/80	0				
				40302	PARAMI	FFAA		

El Salvadoran Case Study

- We consider the following attributes: first name, family name, date of death (year, month, day), and location of death (department, municipality).

- This is an unsupervised application, however, we can test our results on “ground truth” compared to the hand-matched data in departments 1 and 7.



DP prior: Posterior density plot of the number of distinct entities from the Gibbs sampler, along with the posterior mean of 680.66 (black), the true value 680 (red), and 95 percent credible intervals (blue) of [677, 684].

El Salvadoran Case Study

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

Comparison of the PYP, DP, and Uniform priors for the El Salvadoran case study.

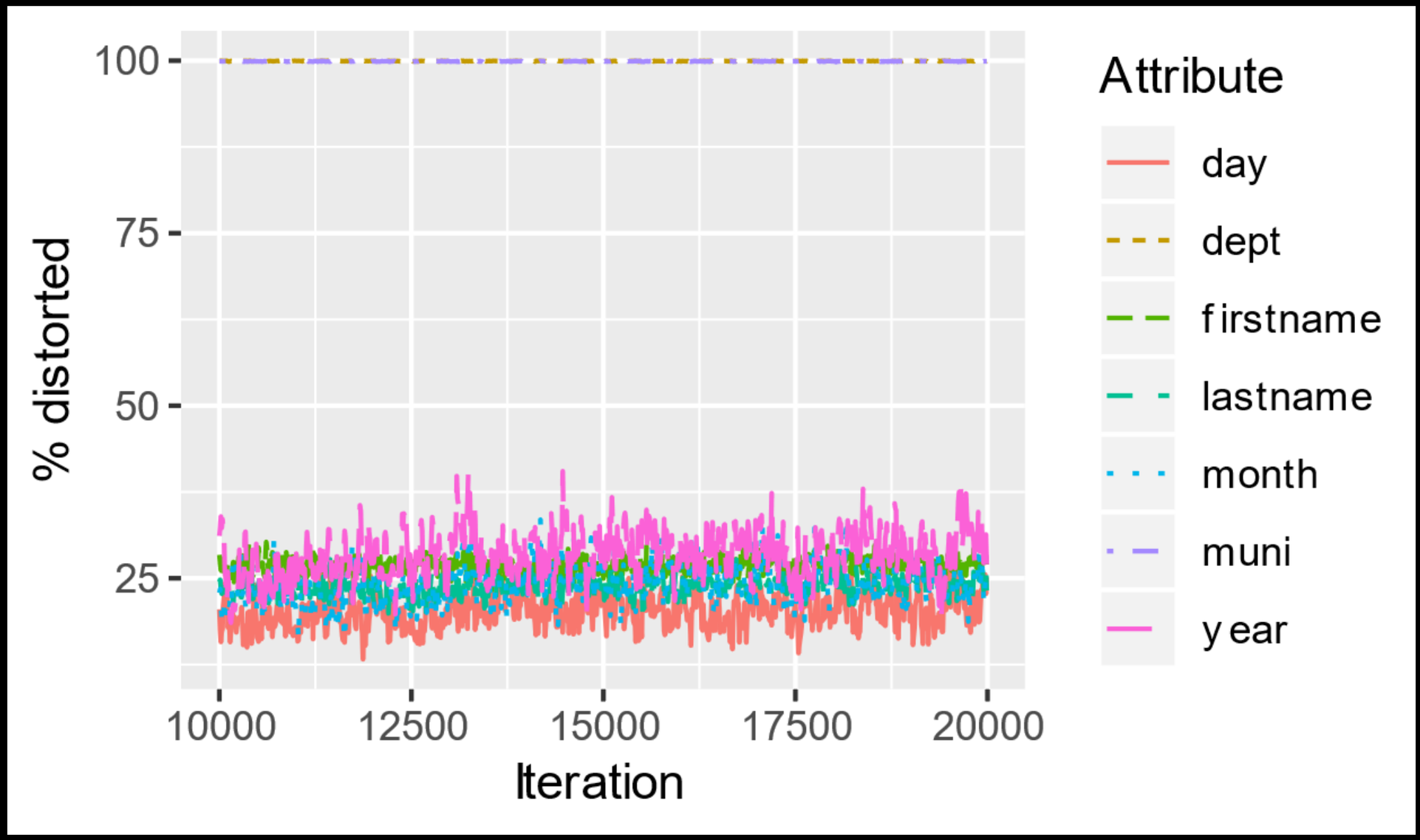
El Salvadoran Case Study

firstname	lastname	year	month	day	dept	muni
ANTONIO	ERNADES	1982	3	21	7	716
MIGUEL	ERNADES	1980	3	22	7	716

firstname	lastname	year	month	day	dept	muni
CARMEN	ALFARO	1982	3	21	7	716
JOSE	ALFARO GAMES	1980	3	22	7	716
CARMEN	ALFARO GAMES	1980	3	22	7	716

Compound names are a major issue so here we have a false positive. The model is overly zealous at linking compound names, such as JOSE ALFARO GAMES and CARMEN ALFARO.

El Salvadoran Case Study



Geographical attributes are 100% distorted.

Thank you!

Questions?

beka@stat.duke.edu

resteorts.github.io