Module X: Bayesian Graphical Entity Resolution

Rebecca C. Steorts

Reading

- Binette and Steorts (2020)
- Steorts, Hall, Fienberg (2016)
- Steorts (2015)

What is “Bayesian”?

1. Setting up a *full probability model* – a joint probability distribution for all observable and unobservable quantities

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\theta}) &- \text{ likelihood} \\
p(\boldsymbol{\theta}) &- \text{ prior}
\end{aligned}
$$

2. Conditioning on observed data – calculating and interpreting the appropriate *posterior distribution*

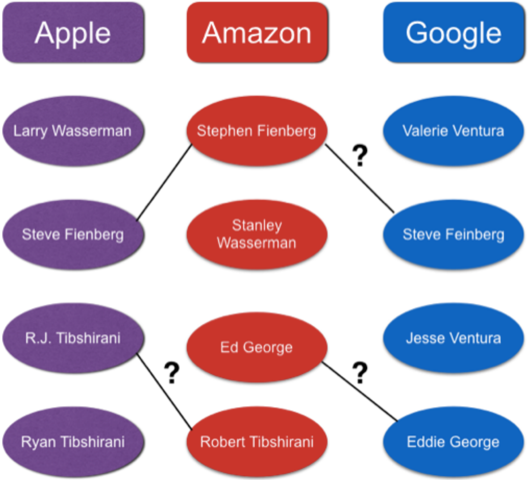
$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Why Bayesian Entity Resolution

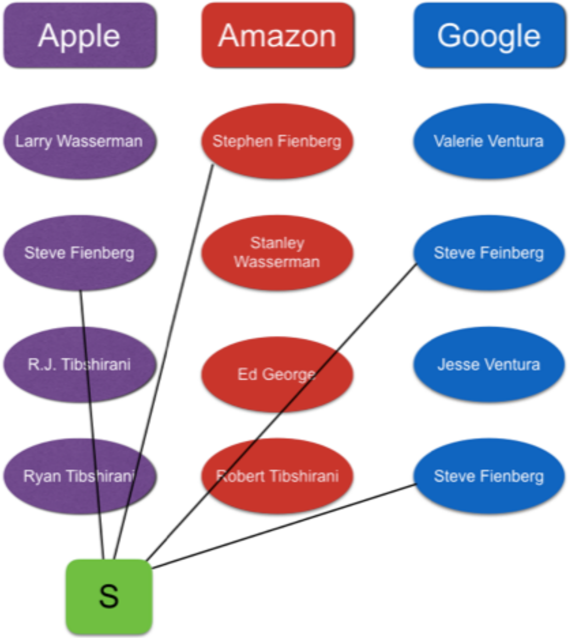
1. Entity resolution can be treated as a clustering problem.
2. Records are clustering to a latent entity.
3. This results in the model becoming a bipartite graph, which allows one to estimate latent individuals across multiple high dimensional databases.
4. The Bayesian paradigm naturally allows uncertainty quantification of the entity resolution process, a full posterior distribution, credible intervals, etc.
5. Theoretical properties have recently been explored for latent variable models, supporting the above approach.

[Copas and Hilton (1990), Tancredi and Liseo (2011), Steorts, Barnes, Neiswanger (2017), Zanella et al. (2016)]

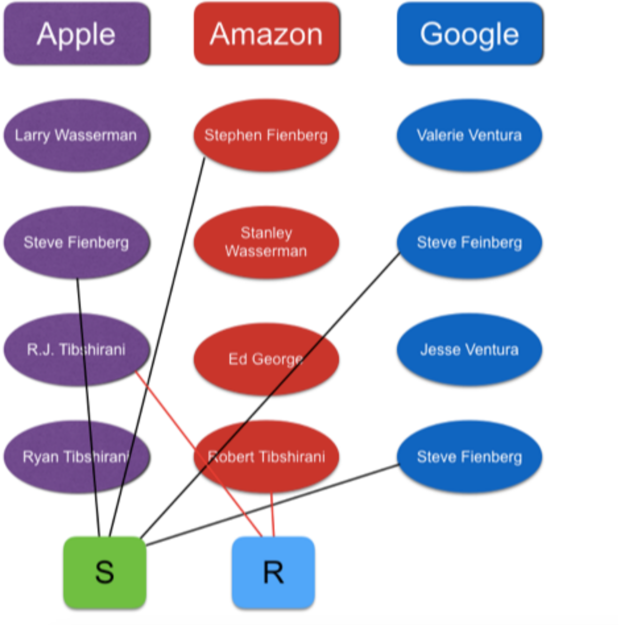
The entity resolution graph



The latent variable approach



The latent variable approach

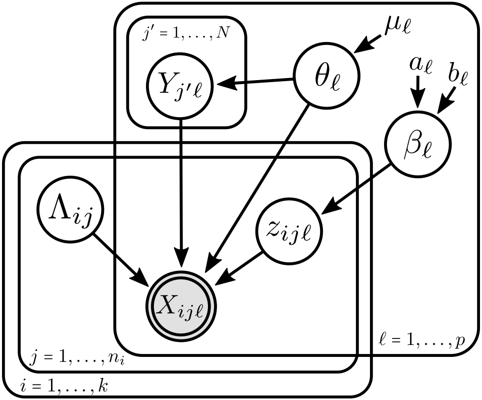


Notation

- $X_{ij\ell}$: observed value of the ℓth field for the jth record in the ith data set, $1 \leq i \leq k$ and $1 \leq j \leq n_i$.
- $Y_{j'\ell}$: true value of the ℓth field for the j'th latent individual.
- λ_{ij}: latent individual to which the jth record in the ith list corresponds. $\mathbf{\Lambda}$ is the collection of these values..
  - e.g. Five records in one list $\mathbf{\Lambda} = \{1, 1, 2, 3, 3\} \rightarrow 3$ latent entities or clusters.
- $z_{ij\ell}$: indicator of whether a distortion has occurred for record field value $X_{ij\ell}$

Graphical Record Linkage

Graphical model representation of Steorts et al. (2016):



- Λ_{ij} represents the linkage structure $\rightarrow$ uniform prior.
- Requires information about the number of latent entities a priori and it is very informative.

Bayesian Entity Resolution

Previous literature: not balanced regarding modeling, handling high-dimensional data, and uncertainty of multiple databases.

- Bayesian model: simultaneously links and de-duplicates.
- Assume records are noisy, distorted.
- We have a novel representation ($\mathbf{\Lambda}$): linkage structure.
- The strength of a Bayesian approach is that transitivity of the linked records is nearly automatic.
- Our data structure provides uncertainty estimates of linked records that can be propagated into later analyses.

[Steorts, Hall, and Fienberg (2014), Steorts, Hall, and Fienberg (2016)]

Empirically Motivated Priors

- The major weakness of Steorts, Hall, and Fienberg (2016) is the fact that it did not handle text (string) data.
- Steorts (2015) overcomes this issue by taking an empirical Bayesian approach, and making extensive comparisons to supervised methods.

Model Specification: String model

- The distortion of string-valued variables is modeled using a probabilistic mechanism based on some measure of distance between the true and distorted strings.

$$P(X_{ij\ell} = w | \lambda_{ij}, Y_{\lambda_{ij}\ell}, z_{ij\ell}) = \frac{\alpha_\ell \exp[-cd(w, Y_{\lambda_{ij}\ell})]}{\sum_{w \in S_l} \alpha_\ell \exp[-cd(w, Y_{\lambda_{ij}\ell})]}$$

where c is a parameter that needs to be specified and d represents a string metric distance e.g. Levenshtein or Jaro-Winkler.

Model Specification: Likelihood Function

$$X_{ij\ell} = w | \lambda_{ij}, Y_{\lambda_{ij}\ell}, z_{ij\ell} \overset{iid}{\sim} \begin{cases} \delta(Y_{\lambda_{ij}\ell}), & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\lambda_{ij}\ell}), & \text{if } z_{ij\ell} = 1 \text{ and } \ell \le p_s \\ G_\ell, & \text{if } z_{ij\ell} = 1 \text{ and } \ell > p_s \end{cases}$$

- $z_{ij\ell} = 0$, then $X_{ij\ell} = Y_{\lambda_{ij\ell}}$
- F_ℓ is the string model in the last slide.
- G_ℓ is the empirical distribution function of the categorical data.

Model Specification: Hierarchical Model

$$
\begin{aligned}
&Y_{\lambda_{ij}\ell} \overset{iid}{\sim} G_\ell \\
&z_{ij\ell}|\beta_{i\ell} \overset{iid}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\
&\beta_{i\ell} \overset{iid}{\sim} \text{Beta}(a, b) \\
&\lambda_{ij} \overset{iid}{\sim} \text{DiscreteUniform}(1, \ldots, \text{N})
\end{aligned}
$$

where a, b, N are unknown parameters that must be estimated or fixed.

- $\beta_{i\ell}$ represent the distortion probabilities of the fields.
- The parameters a and b for the Beta prior need to be specified.
- The number of latent entities or clusters needs to be specified in advance.

blink package

R package that removes duplicate entries from multiple databases using the empirical Bayes graphical method:

```r
library("blink")
```

- Formatting data for use with blink
- Tuning parameters
- Running the Gibbs sampler (estimate model parameters)
- Output

RLdata500 data

We will continue with the RLdata500 dataset in the blink package consisting of 500 records with 10% duplication.

```r
library(RecordLinkage)
```

```
## Loading required package: DBI

## Loading required package: RSQLite

## Loading required package: ff

## Loading required package: bit

##
## Attaching package: 'bit'

## The following object is masked from 'package:base':
##
##     xor

## Attaching package ff
```