Module X: Bayesian Graphical Entity Resolution

Rebecca C. Steorts

Reading

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

What is "Bayesian"?

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

$$p(\boldsymbol{x}|\boldsymbol{\theta}) - \text{ likelihood}$$
$$p(\boldsymbol{\theta}) - \text{ prior}$$

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

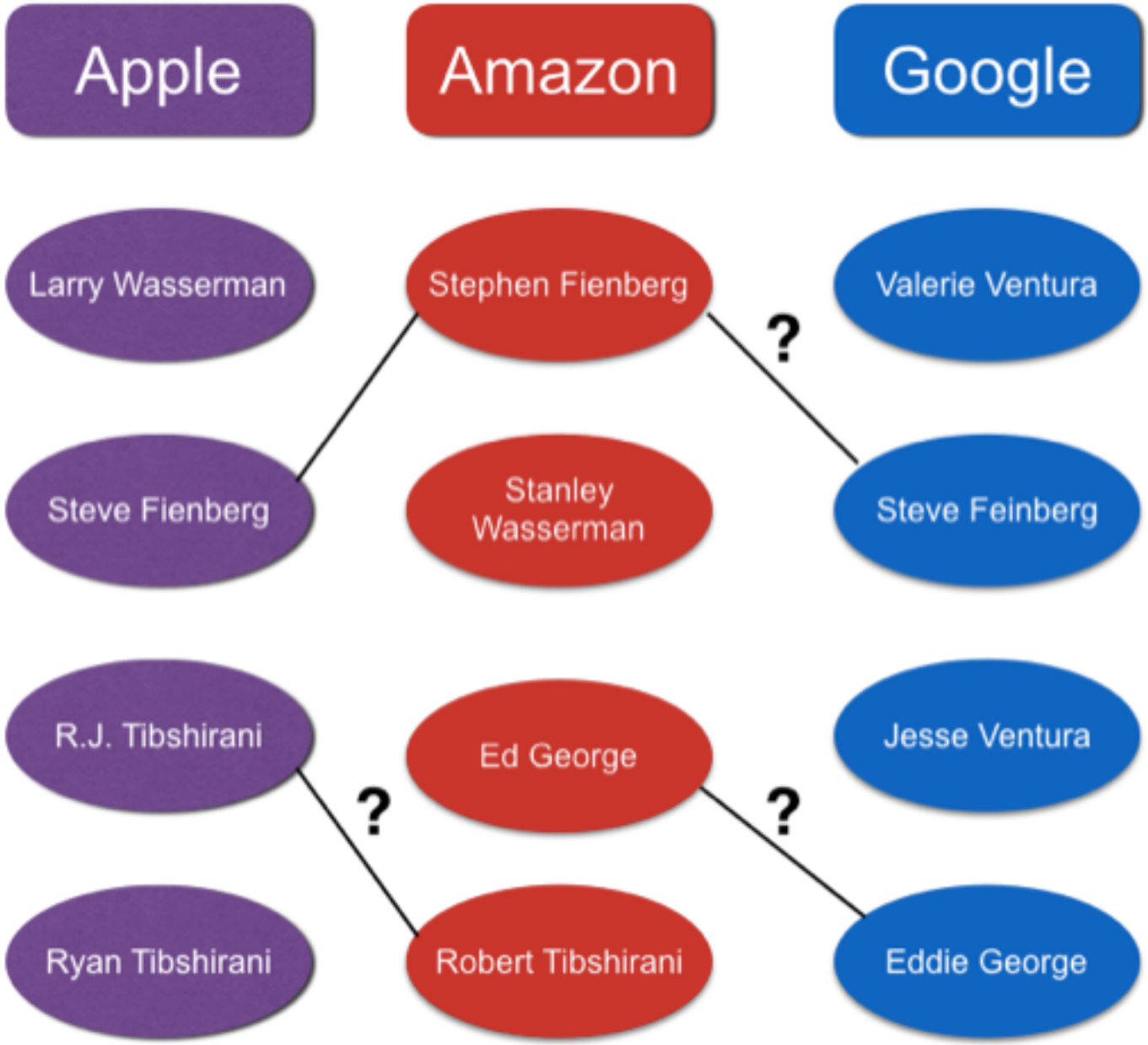
$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Why Bayesian Entity Resolution

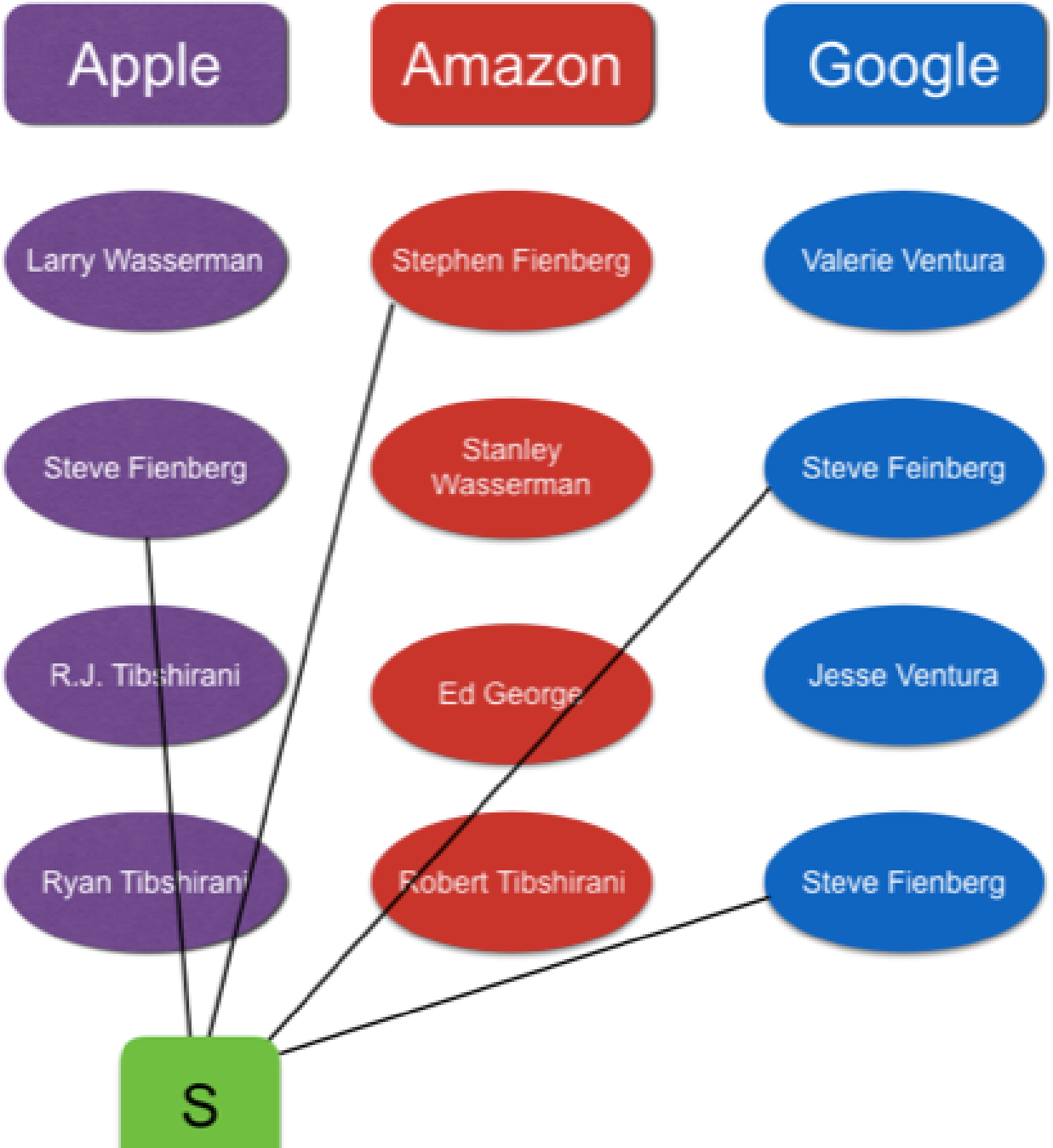
1. Entity resolution can be treated as a clustering problem.
2. Records are clustering to a latent entity.
3. This results in the model becoming a bipartite graph, which allows one to estimate latent individuals across multiple high dimensional databases.
4. The Bayesian paradigm naturally allows uncertainty 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)]

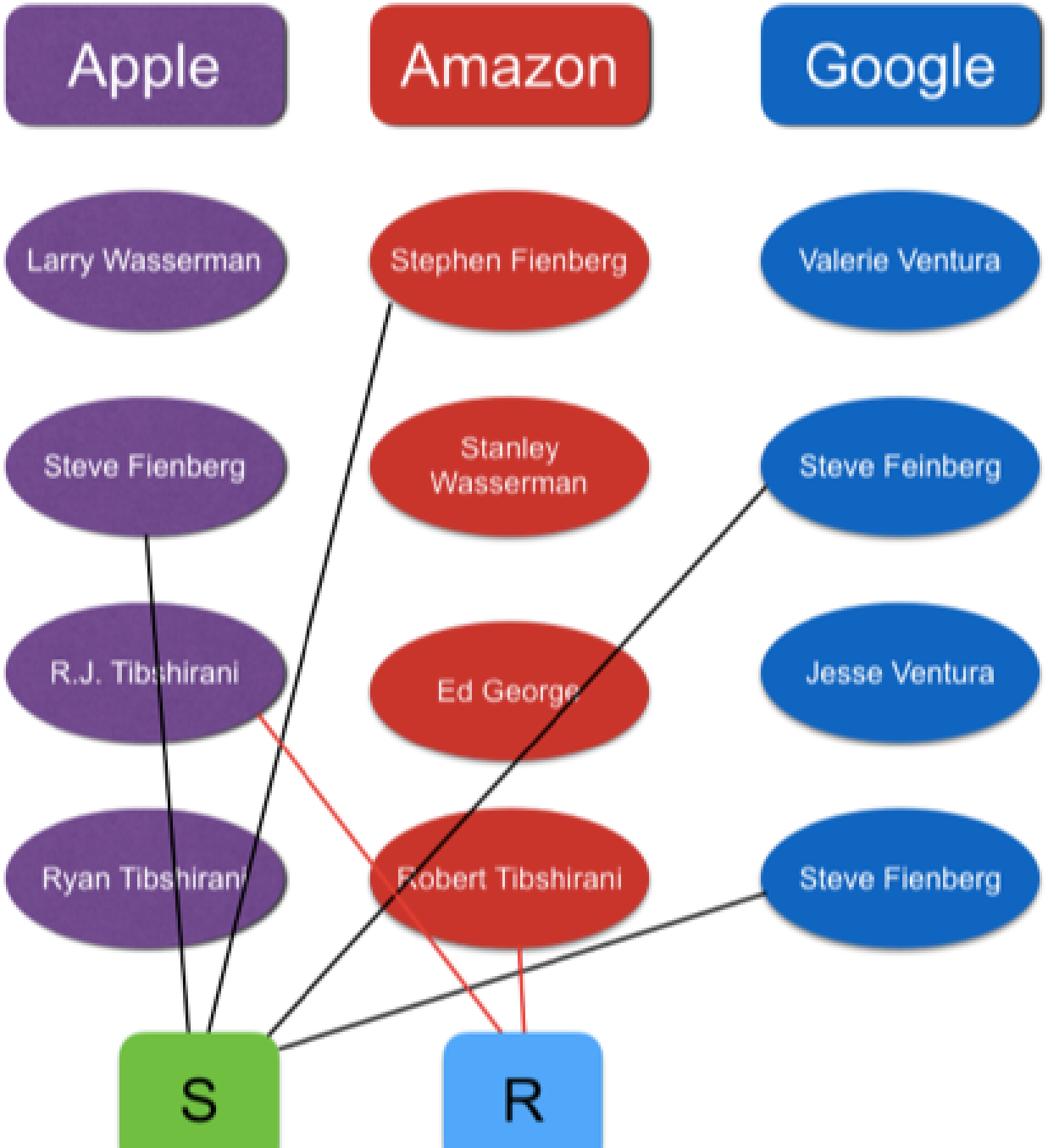
The entity resolution graph



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 $X_{ij\ell}$