

# Introduction to Bayesian record linkage

Brenda Betancourt and Andee Kaplan

Duke University  
Department of Statistical Science

February 8, 2018

Slides available at <http://bit.ly/cimat-bayes>

# Why Bayes?

A Bayesian framework is suitable to solve the following problems:

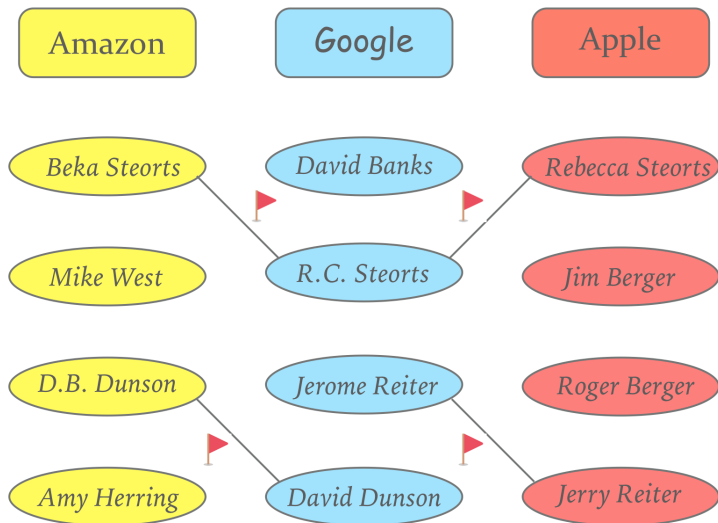
- Exact computation of the probability that each pair is a match, conditional on the observed data.
  - Results conditioning on observed events are more directly interpretable than those obtained by conditioning on unobservable hypotheses.
- Propagating linkage error as an added component of uncertainty in the estimation process.
  - Relevant for subsequent modeling.

# Clustering Approaches

- Record linkage can be naturally seen as a clustering problem.
  - Supervised and unsupervised approaches.
- Records representing the same individual are clustered to a latent entity producing a partition of the data.
  - Steorts, R., Hall, R., and Fienberg, S.E. (2016). A Bayesian Approach to Graphical Record Linkage and De-duplication, Journal of the American Statistical Association, 111:516 (1660-1672).
  - Sadinle, M. (2014). Detecting duplicates in a homicide registry using a bayesian partitioning approach. The Annals of Applied Statistics, Vol. 8, No. 4, 2404–2434

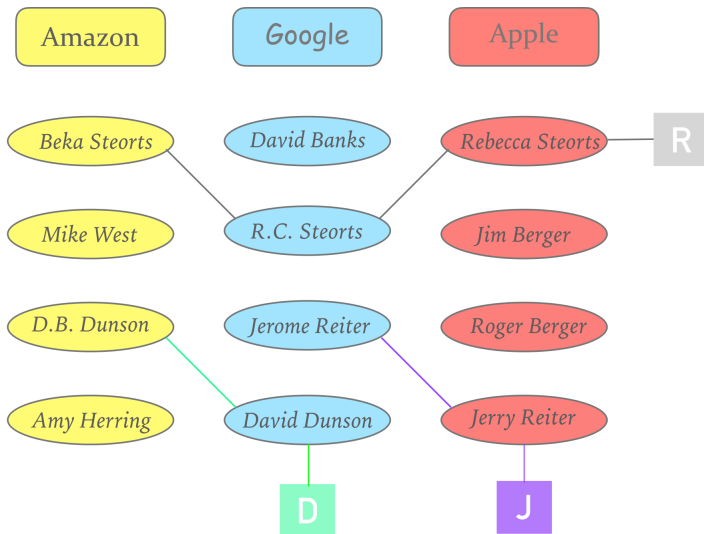
# Record Linkage and Clustering

Which records correspond to the same person?



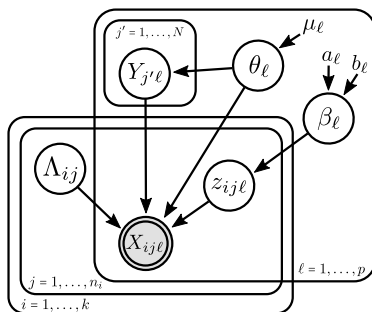
# Record Linkage and Clustering

Each entity is associated with one or more records and the goal is to recover the latent entities (clusters).



# Graphical Record Linkage

Graphical model representation of [Steorts et al. \(2016\)](#):



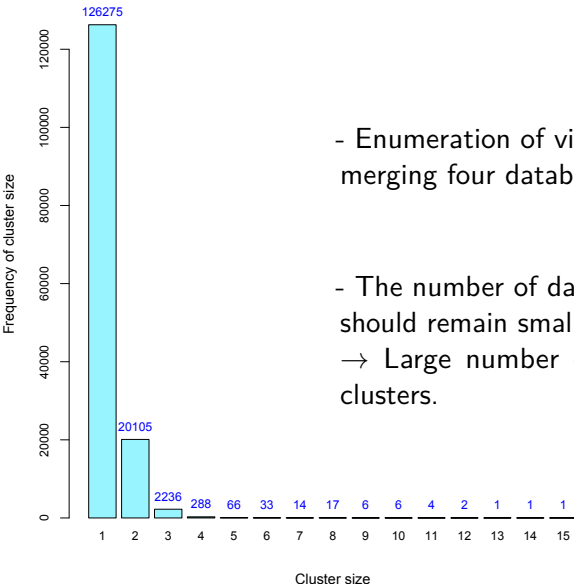
- $\Lambda_{ij}$  represents the linkage structure  $\rightarrow$  **uniform prior**.
- Requires information about the number of latent entities a priori and it is very informative.

# Partition-based Bayesian clustering models

Goal: cluster  $N$  data points  $x_1, \dots, x_N$  into  $K$  clusters.

- Place a prior distribution over partitions of  $[N] = \{1, \dots, N\}$
- Let  $C_N$  be a random partition of  $[N]$
- $C_N$  represented by a set of cluster assignments  $z_1, \dots, z_N$ .
- The number of clusters  $K$  does not need to be specified a priori  
→ **Non-parametric** latent variable approach.

# Record Linkage and Microclustering



- Enumeration of victims of killings in Syria merging four databases.

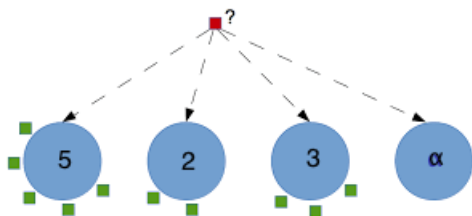
- The number of data points in each cluster should remain small even for large data sets  
→ Large number of singletons and small clusters.



# Mixture Models

Other clustering tasks require models that assume cluster sizes grow linearly with the size of the data set.

- Dirichlet process (DP)  $\implies$  Chinese Restaurant Process (CRP)



- Carmona C., Nieto-Barajas L., Canale A. (2017), Model-based approach for household clustering with mixed scale variables <https://arxiv.org/abs/1612.00083>.

# Microclustering models

- Prior distributions on partitions that are suitable for the microclustering problem.
  - Zanella et al (2016). Flexible Models for Microclustering with Applications to Entity Resolution, Advances in Neural Information Processing Systems (NIPS), Vol. 29, pp 1417-1425.
- Scalable sampling algorithm in combination with blocking techniques.
  - Miller et al (2015). The Microclustering Problem: When the Cluster Sizes Don't Grow with the Number of Data Points. NIPS Bayesian Nonparametrics: The Next Generation Workshop Series.

# blink

Describe in detail the blink model, likelihood, prior and hyperparameters

# blink package

## Example (RL500)