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.

The Fellegi-Sunter approach (1969)

- Represent every pair of records using vector of features that describe similarity between individual record fields.
  - Use string metrics (Jaro-Winkler) and edit-distances for names and strings of numbers.

- Place feature vectors for record pairs into three classes: matches (M), nonmatches (U), and possible matches.

- Let $P(\gamma|M)$ and $P(\gamma|U)$ be probabilities of observing a feature vector γ for a matched and nonmatched pair, respectively.

The Fellegi-Sunter approach (1969)

- Perform record-pair classification by calculating the ratio $(P(\gamma|M)/P(\gamma|U))$ for each candidate record pair.
- Establish two thresholds based on desired error levels to optimally separate the ratio values for matches, possibly matches, and nonmatches.
- Drawbacks: only for two files, no trasitive closures.

Fellegi-Sunter generalization

- Extension for multiple files solving the problem of non–transitive decisions.
- Provide matching probabilities for the record K-tuples, necessary to incorporate the uncertainty of the linkage procedure in posterior analysis
- M. Sadinle and S.E. Fienberg (2013). "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." J. Amer. Statist. Assoc., 108 (502), 385–397.

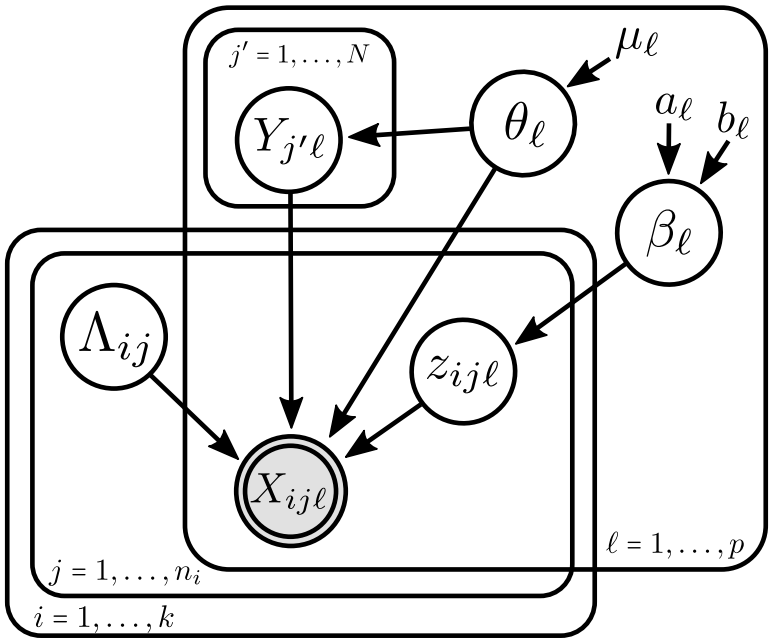
Clustering Approaches

- Record linkage can be naturally seen as a clustering problem.
  - Supervised and unsupervised approches.

- 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 275 bayesian partitioning approach. The Annals of Applied Statistics, Vol. 8, No. 4, 2404–2434

Graphical Record Linkage

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



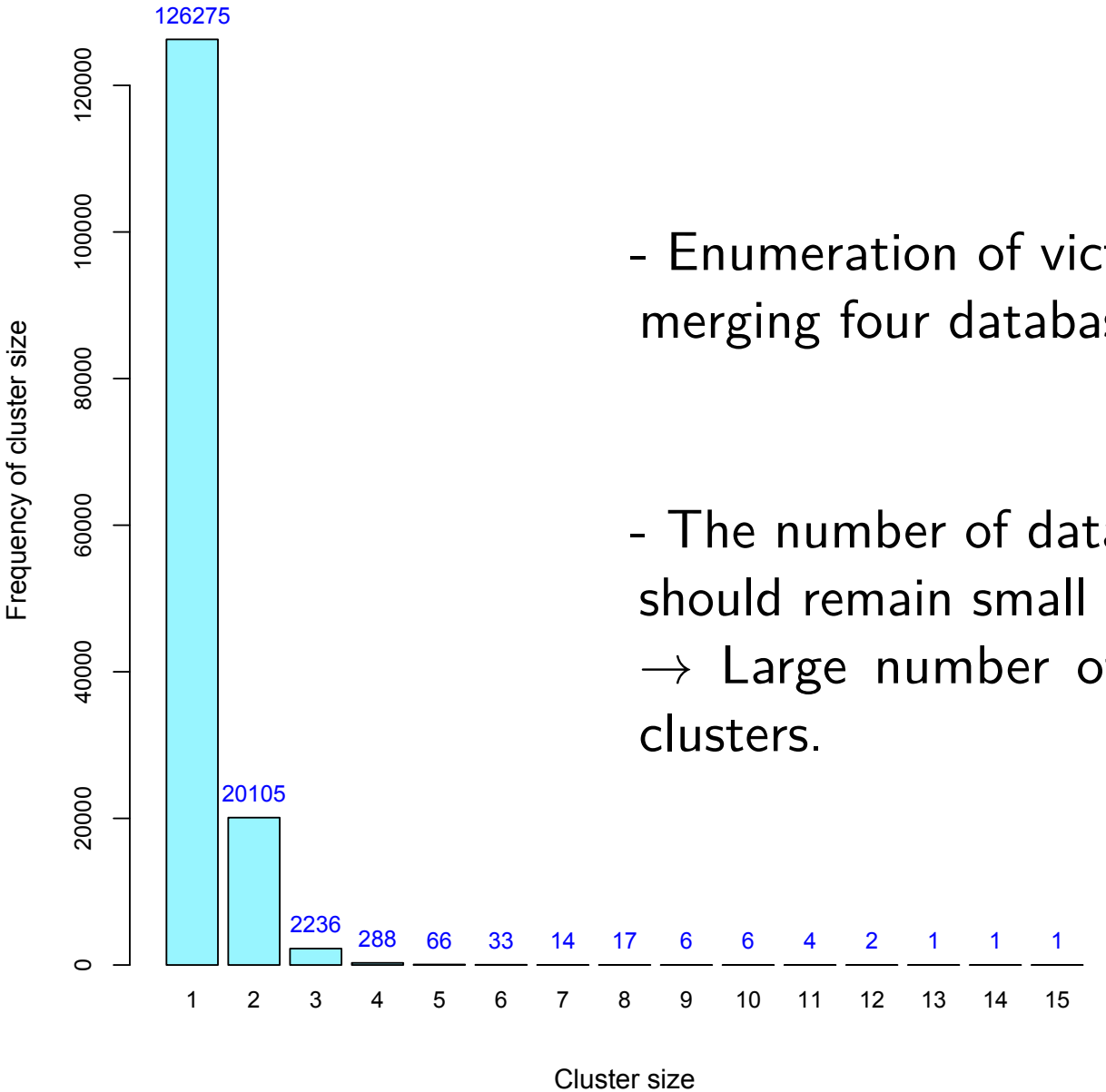
- Λ_{ij} represents the linkage structure $\rightarrow$ uniform prior.
- Requires information about the number of latent entities a priori and it is very informative. [[Ask Beka for Fig 2 in SMERED paper, network representation]]

Partition-based Bayesian clustering models

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

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

Record Linkage and Microclustering



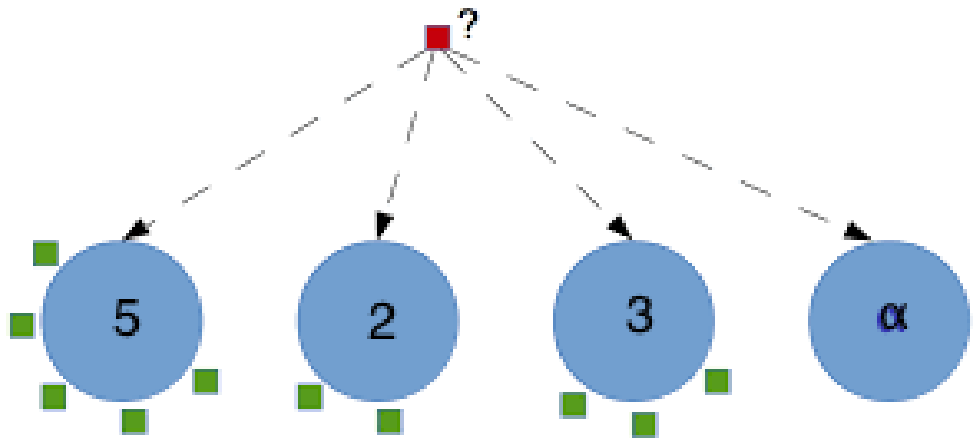
- Enumeration of victims of killings in Syria merging four databases.

- The number of data points in each cluster should remain small even for large data sets $\rightarrow$ 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) $\Longrightarrow$ 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 deatil the blink model, likelihood , prior and hyperparameters

blink package

Example (RL500)