Why infinite exchangeable mixture models fail for "sparse" data sets yet microclustering succeeds

Rebecca C. Steorts

Department of Statistical Science, Duke University
affiliated faculty in Computer Science, Biostatistics, the information initiative at Duke (iid), and the Social Science Research Initiative (SSRI)

joint work with Jeff Miller, Brenda Betancourt, Abbas Zaidi, and Hanna Wallach, Giacomo Zanella

April 6, 2016

Computational social science

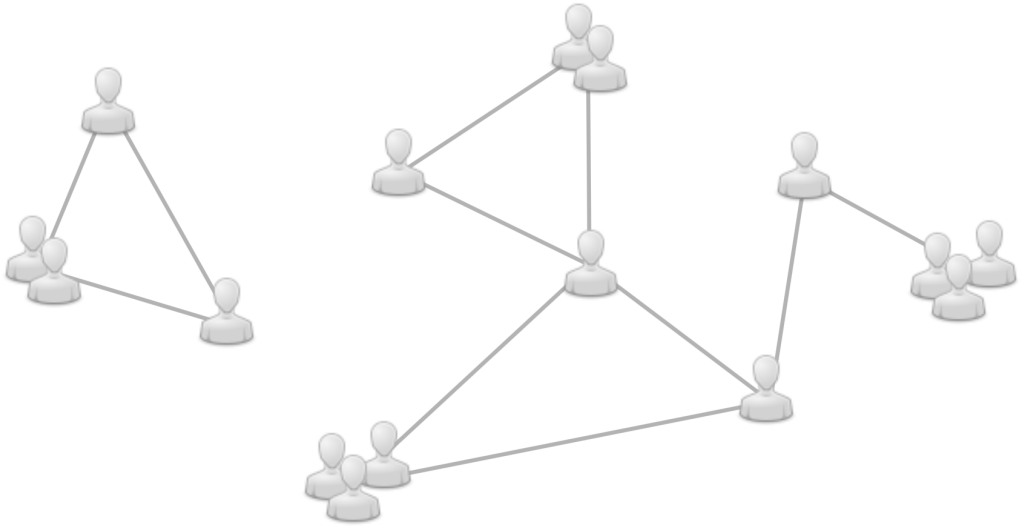
"Computational social science is the study of social phenomena using digitized information and streaming data along with computational and statistical methods."

-based on definition from Modeling Topic-Partitioned Network Structure, Hanna Wallach, NIPS, 2015

Social processes, entities, and clusters

[picture from Modeling Topic-Partitioned Network Structure, Hanna Wallach, NIPS, 2015]

Structure

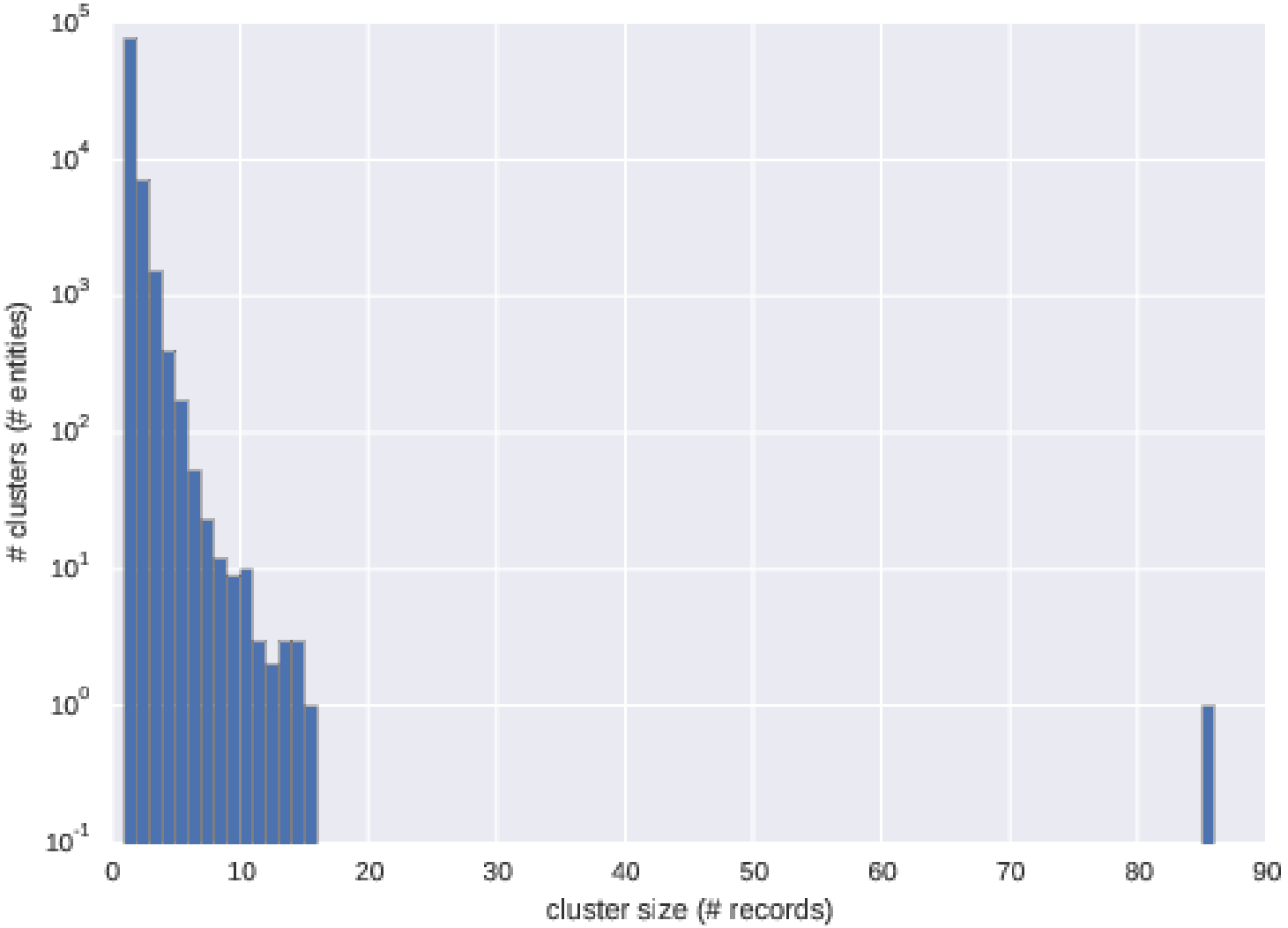


[picture from Modeling Topic-Partitioned Network Structure, Hanna Wallach, NIPS, 2015]

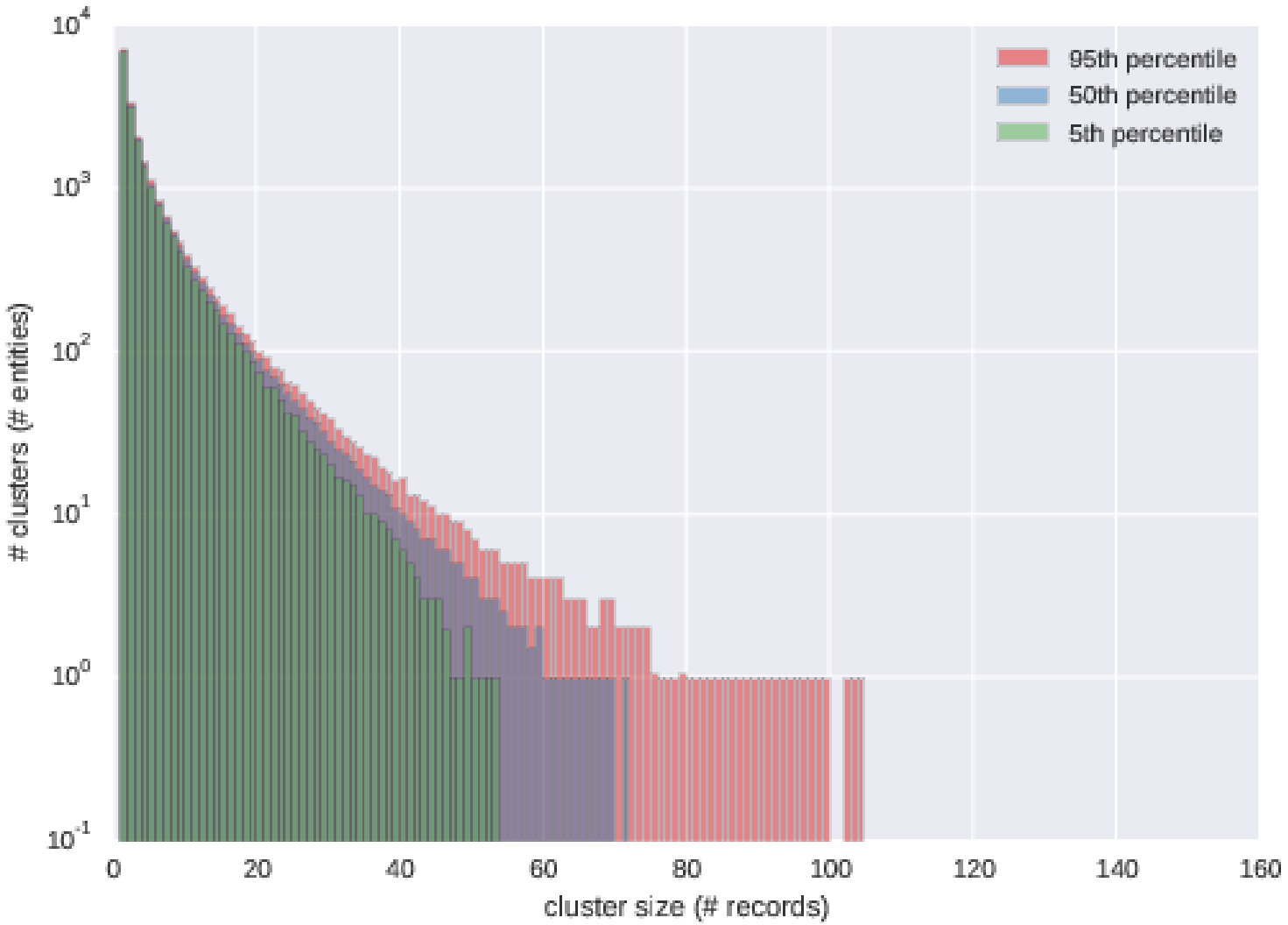
Clustering

- Popular clustering approaches have limitations.
- Number of data points per cluster is not expected to grow without bound.
- Propose new clustering model(s) that captures this behavior.
- Applications: Medical data, official statistics, author disambiguation, investigative journalism, human rights violations, customer and transaction records, credit reports, and others.

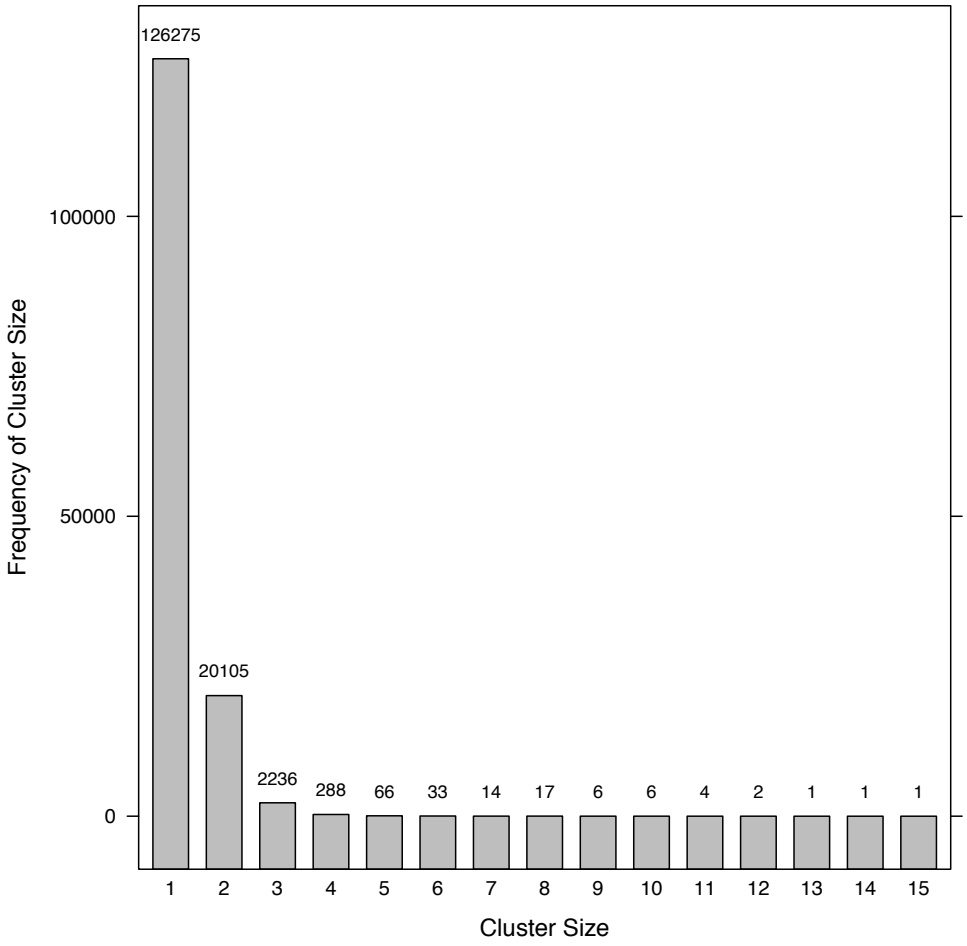
Motivation: Campaign Finance Data (100K records)



Chinese Restaurant Process anyone?



Second Motivation: Syrian Conflict



Problem of interest: what is the death toll for the Syrian conflict?
Data: Four human rights data sets with duplicated death records.

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

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

Classic examples are the Dirichlet process (DP) and the Chinese Restaurant Process (CRP).

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

Classic examples are the Dirichlet process (DP) and the Chinese Restaurant Process (CRP).

More generally, we think of all infinite mixture models (Pitmor-Yor Process (PYP) and the Kingman Paintbox).

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

Classic examples are the Dirichlet process (DP) and the Chinese Restaurant Process (CRP).

More generally, we think of all infinite mixture models (Pitmor-Yor Process (PYP) and the Kingman Paintbox).

We review infinite mixture models and contrast these with our approach.

Clusters and Partitions

To cluster N data points $x_1, \ldots, x_N$ using a partition-based Bayesian clustering model, one first places a prior over partitions of $[N] = \{1, \ldots, N\}$.

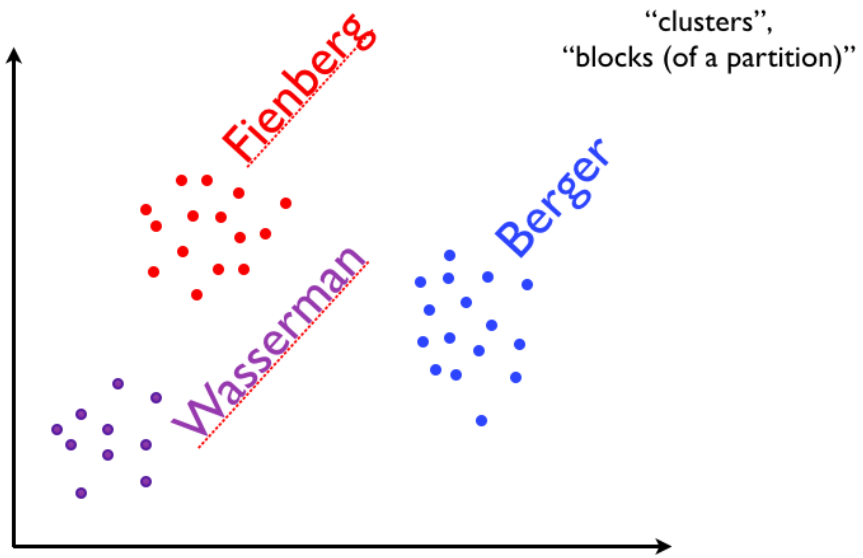
Let C_N be a random partition of $[N]$.

Clusters and Partitions

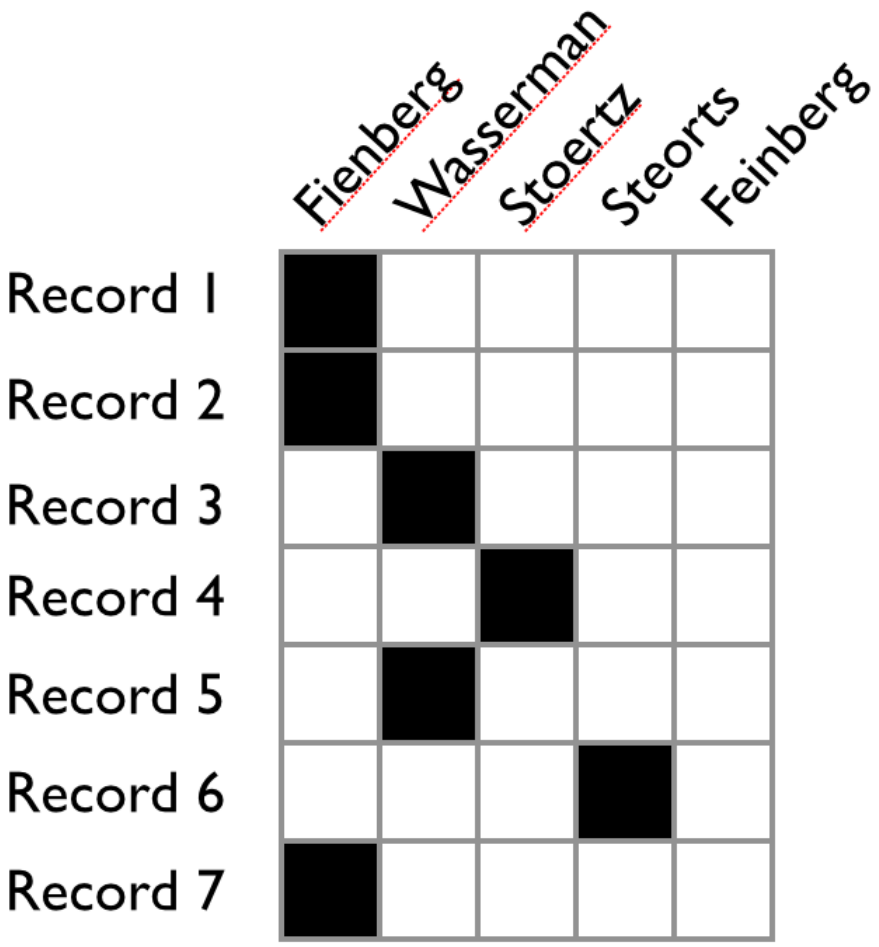
To cluster N data points $x_1, \ldots, x_N$ using a partition-based Bayesian clustering model, one first places a prior over partitions of $[N] = \{1, \ldots, N\}$.

Let C_N be a random partition of $[N]$.

C_N is implicitly represented by a set of cluster assignments $z_1, \ldots, z_N$.



Clusters and Partitions



- No ordering imposed here.
- Also, records can only belong to one cluster assignment. (Don't confuse this with topic modeling).

Infinite mixture model

One regards these cluster assignments as the first N elements of an infinite sequence $z_1, z_2, \ldots$, drawn a priori from

$$\boldsymbol{\pi} \sim H \quad \text{and} \quad z_1, z_2, \ldots \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \boldsymbol{\pi}, \tag{0.1}$$

where

- H is a prior over $\boldsymbol{\pi}$ and
- $\boldsymbol{\pi}$ is a vector of mixture weights with

$$\sum_{\ell} \pi_\ell = 1 \text{ and } \pi_\ell \geq 0$$

for all ℓ.

Common mixture models

1. Finite mixtures where the dimensionality of π is fixed and H is usually a Dirichlet distribution;
2. Finite mixtures where the dimensionality of π is a random variable (Richardson and Green, 1997; Miller and Harrison, 2015),
3. Dirichlet process (DP) mixtures where the dimensionality of π is infinite (Sethuraman, 1994),
4. Pitman Yor process (PYP) mixtures, which generalize DP mixtures (Ishwaran and James, 2003).

The Kingman Paintbox

The Kingman Paintbox

For clustering, the de Finetti mixing measure that gives rise to exchangeability is the Kingman paintbox.

As we will see, Kingman will not be suitable to record linkage tasks.

Microclustering

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if M_N is $o_p(N)$, where M_N is the size of the largest cluster in C_N.

Microclustering

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if M_N is $o_p(N)$, where M_N is the size of the largest cluster in C_N.

A clustering model exhibits the microclustering property if C_N implied by that model satisfies the above definition.

Microclustering and Mixture Models

- No mixture model can exhibit the microclustering property (unless its parameters are allowed to vary with N).
- Kingman's paintbox theorem implies that any exchangeable partition of $\mathbb{N}$ is
  1. either equal to the trivial partition in which each part contains one element or
  2. satisfies $\liminf_{N\to\infty} M_N \,/\, N > 0$ with positive probability.

Microclustering and Mixture Models (continued)

- By Kolmogorov's extension theorem, a sequence of random partitions $(C_N : N = 1, 2, \ldots)$ corresponds to an exchangeable random partition of $\mathbb{N}$ whenever
  (a) each C_N is exchangeable and
  (b) the sequence is consistent in distribution[1]
- Therefore, to obtain a nontrivial model that exhibits the microclustering property, one must sacrifice either (a) or (b).
- Wallach et al. (2010) sacrificed (a). We instead sacrifice (b).

[1] i.e., if $N' < N$, the distribution of $C_{N'}$ coincides with the marginal of $C_{N'}$ obtained using the distribution of C_N.

Our Approach

- Propose a model that satisfies the microclustering property.
- Seek flexible and robust models.
- New sampling algorithm.
- Initial results on simulated and real data with comparisons to infinite mixture models.

[Miller, Betancourt, Zaidi, Wallach, and Steorts (2015), `http://arxiv.org/abs/1512.00792`]

A Microclustering Model

Let K be the potential number of clusters. Define $N = \sum_{k=1}^{K} N_k$.

A Microclustering Model

Let K be the potential number of clusters. Define $N = \sum_{k=1}^{K} N_k$.

$$K \sim \mathrm{NegBin}\,(a, q) \quad \text{and} \quad N_1, \ldots, N_K \mid K \overset{\text{iid}}{\sim} \mathrm{NegBin}\,(r, p),$$

for $a, r > 0$ and $q, p \in (0, 1)$.

A Microclustering Model

Let K be the potential number of clusters. Define $N = \sum_{k=1}^{K} N_k$.

$$K \sim \mathrm{NegBin}\,(a, q) \quad \text{and} \quad N_1, \ldots, N_K \mid K \overset{\text{iid}}{\sim} \mathrm{NegBin}\,(r, p),$$

for $a, r > 0$ and $q, p \in (0, 1)$.

Given $N_1, \ldots, N_K$, generate a set of cluster assignments $z_1, \ldots, z_N$ by drawing a vector uniformly at random from the set of permutations of

$$(\underbrace{1, \ldots, 1}_{N_1 \text{ times}}, \underbrace{2, \ldots, 2}_{N_2 \text{ times}}, \ldots\ldots, \underbrace{K, \ldots, K}_{N_K \text{ times}}).$$

The Marginal Distribution

Marginal distribution of C_N leads to a type of reseating algorithm.

A Reseating Algorithm

Consider $P(C_N \mid N, C_N \backslash n)$, where $C_N \backslash n$ is the partition obtained by removing element n from C_N:

- for $n = 1, \ldots, N$, reassign element n to
  - an existing cluster $c \in C_N \backslash n$ with probability $\propto |c| + r$,
  - a new cluster with probability $\propto (|C_N \backslash n| + a)\, \beta r$.

A Reseating Algorithm

Consider $P(C_N \mid N, C_N \backslash n)$, where $C_N \backslash n$ is the partition obtained by removing element n from C_N:

- for $n = 1, \ldots, N$, reassign element n to
  - an existing cluster $c \in C_N \backslash n$ with probability $\propto |c| + r$,
  - a new cluster with probability $\propto (|C_N \backslash n| + a)\, \beta r$.

There isn't a richer get richer property induced!

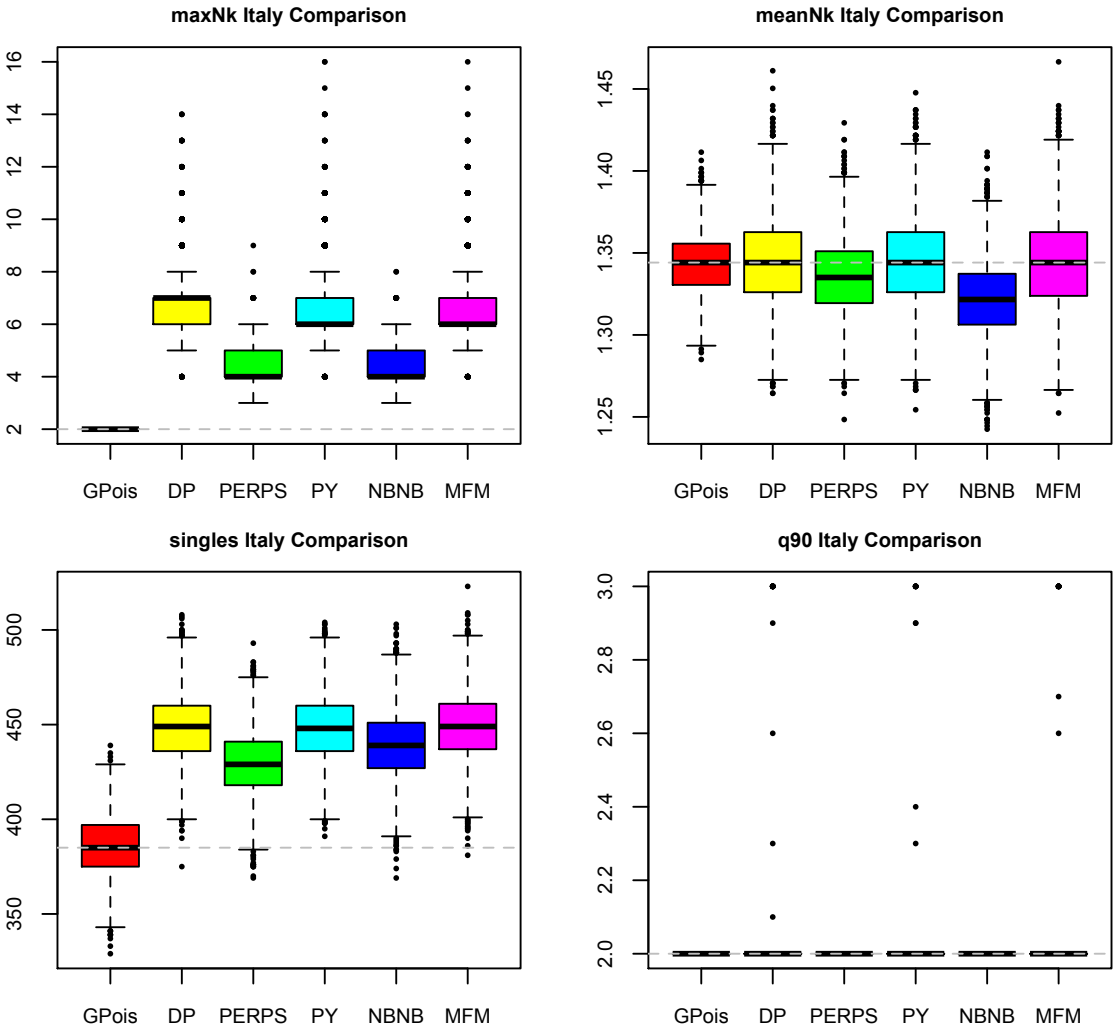
Preliminary Results

- Compare our methods models to several commonly used infinitely exchangeable clustering models: mixtures of finite mixtures MFM, DP mixtures, and PYP mixtures.
- We assess how well each model "fits" partitions using prior predictive checks.
- The prior predictive checks are evaluated at the MLE for the various models.
- We assess "fit" based on "summary statistics," e.g, singletons, quantiles, etc.

The Data

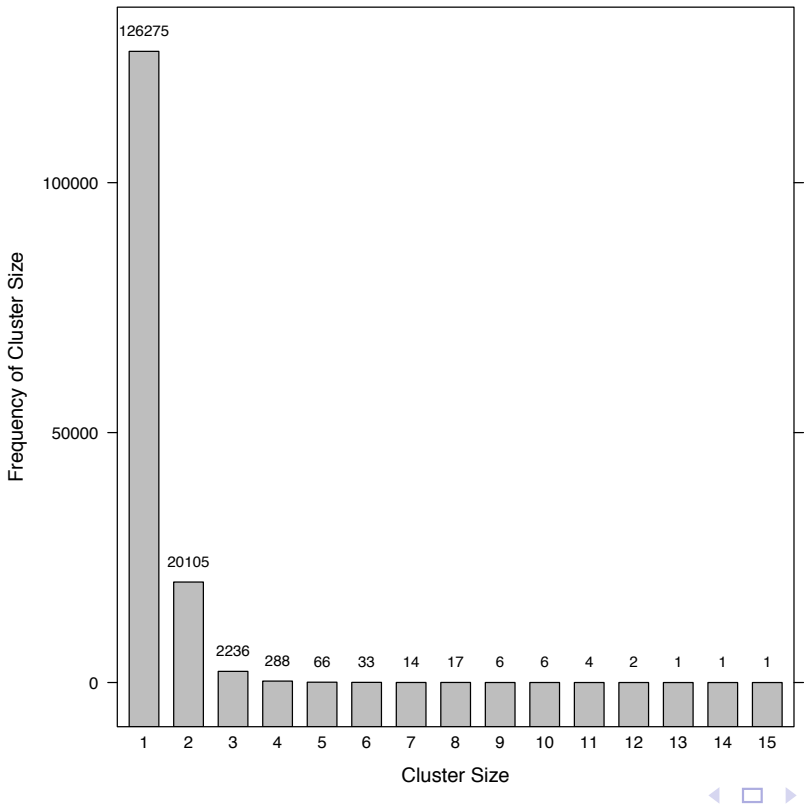
- Survey data from Italy.
- 74% of data are singletons.
- Ground truth from social security numbers.

Prior predictive checks



Why do we do better?

1. Theoretically, we're handling the problem correctly. (Using an infinite process is "cheating.")
2. We do better than the DPMs and MFMs in situations when the data has lots of singletons.
3. Why does the GPois do better?



How does one do inference?

In real world clustering tasks, data points $x_1, \ldots, x_n$ and N are observed.

C_N is latent.

Recall $|C_N|$ denotes the size of the partition C_N, which may be smaller than K. Hence, the number of entities is denoted by $|C_N|$.

For illustration, we assume the clustering task is record linkage.

Inference for record linkage

Assume records $x_{\ell n}$ where:

- ℓ indexes the features within a record (gender, DOB, etc)
- n indexes how many records we observe.

Let $\zeta : \bigcup_{N=0}^{\infty}(\mathcal{C}_N \times [N]) \to \{1, 2, \ldots\}$ be a function that maps a partition C_N and a record n to its latent cluster assignment.[2] Then

$$C_N \sim \text{NBNB}(a, q, r, p) \tag{0.2}$$

$$z_n \mid C_N = \zeta(C_N, n) \tag{0.3}$$

$$x_{\ell,n} \sim \text{DirichletMultinomial}(\delta_\ell), \tag{0.4}$$

where δ_ℓ is assumed known. When $m(x_c)$ can easily be computed for any element c, then we can easily sample from the posterior on partitions, $p(C_N \mid x_1, \ldots, x_n)$

[2] For example, $\zeta(\{\{1,3,4\},\{2,5\},\{6\}\},4) = 1$, $\zeta(\{\{1,3,4\},\{2,5\},\{6\}\},6) = 3$ because in this partition, record 4 is in cluster 1, while record 6 is in cluster 3.

Inference on Italian Survey Data

We return to the Italian survey data

We work with 9 categorical variables about income and wealth.

- We do comparisons with other models.
- We look also at the FNR and FDR.

Table: Results with database with 9 variables. The true number of clusters is 587.

Prior	$\hat{K}$	SD	FNR	FDR
DP	609.4	3.25	0.371	0.471
PYP	621.5	2.78	0.376	0.408
PERPS	606.5	3.13	0.381	0.416
NBNB	620.1	2.33	0.376	0.411

All models overestimate K.

DP and PERPS do the best in terms of a best estimate with error rates.

NBNB and PYP do slightly worse in terms of inference.

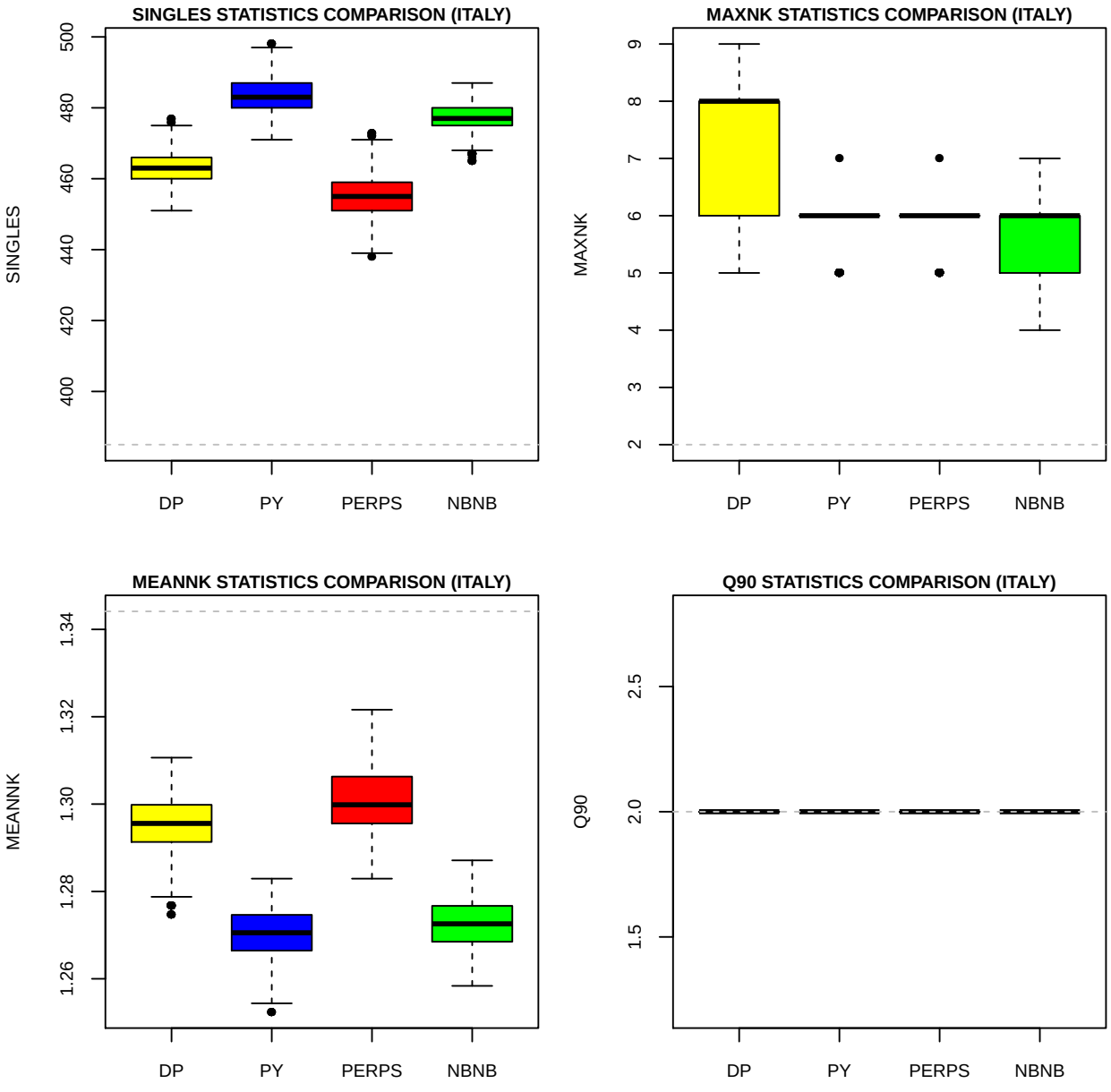


Figure: Results with database with 9 variables. The dashed line represents the true statistics value.

Reseating algorithms

In general, these reseating algorithms are quite slow (due to partitioning) and may be slow to mix!

Propose a solution to this that is similar in spirit to Jain and Neal (2007).

You need a chaperone (actually two)!

- Let C_N denote a partition of $[N]$ and let $x_1, \ldots, x_N$ denote the N observed data points.
- If we let $c_n \in C_N$ denote the cluster containing element n, then each iteration consists of:

1. Randomly choose two *chaperones*, $i, j \in \{1, \ldots, N\}$ from a distribution $P(i, j \mid x_1, \ldots, x_N)$ where the probability of i and j given $x_1, \ldots, x_N$ is greater than zero for all $i \neq j$.
   - This distribution must be independent of the current state of the Markov chain C_N; however, crucially, it may depend on the observed data points $x_1, \ldots, x_N$.
2. Reassign each $n \in c_i \cup c_j$ by sampling from $P(C_N \mid N, C_N \backslash n, c_i \cup c_j, x_1, \ldots, x_N)$.

What do these moves look like?

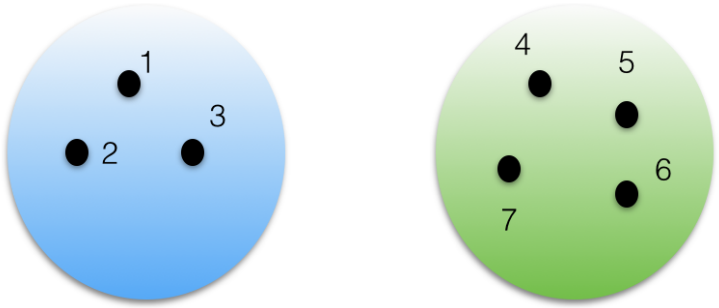


Figure: Two clusters: C_1 and C_2 with assigned elements 1,2,3 and 4,5,6,7 respectively.

What do these moves look like?

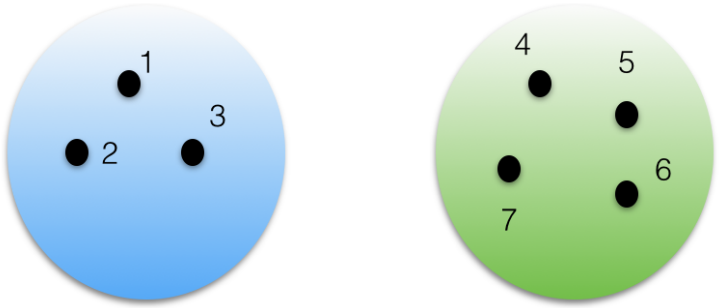
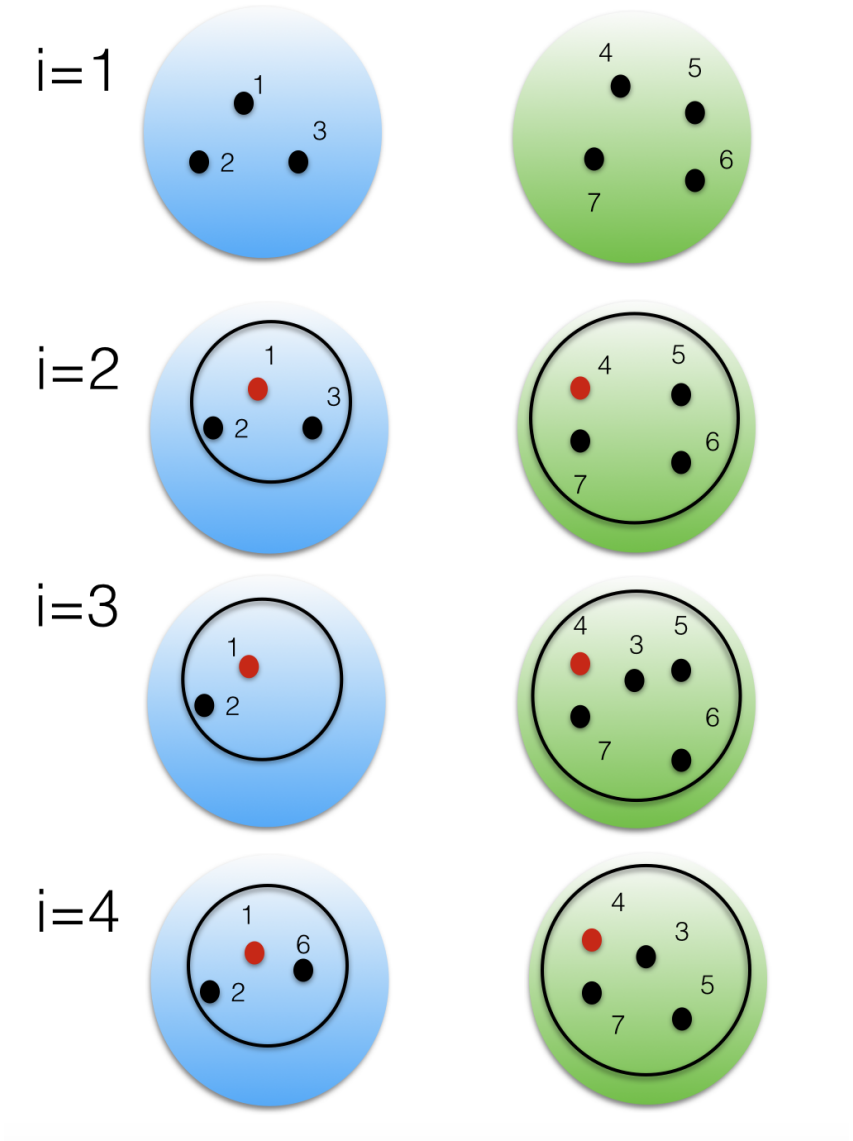


Figure: Two clusters: C_1 and C_2 with assigned elements 1,2,3 and 4,5,6,7 respectively.

Four toy iterations of "chaperones"



Properties

1. The algorithm is geometrically ergodic.
2. Not restricted to uniform moves.
3. This "should" help with better mixing.
4. Understanding how to pick good restricted Gibbs moves is in progress.

Coming soon to a conference near you...

- Inference for all models with comparisons.
- Scalable MCMC using chaperones approach.
- Applying this to data from the Syria conflict (duplicated deaths).

Ongoing work with Jeff Miller, Brenda Betancourt, Abbas Zaidi (Duke University), Hanna Wallach (MSR and UMass Amherst) and Giacomo Zanella (University of Warick).

Thank you!
Questions?
beka@stat.duke.edu

Thank you to the John Templeton Foundation (Metaknowlege Network) and to NSF SES 1534412 for support of this research.
Disclaimer: This work is the view point of the researchers alone and not the funding agencies/foundations.