

# Flexible Models for Microclustering with Application to Entity Resolution

Brenda Betancourt,<sup>1</sup> Giacomo Zanella,<sup>2</sup> Hanna Wallach,<sup>3</sup> Jeffrey W. Miller,<sup>4</sup> Abbas Zaidi,<sup>1</sup> and Rebecca C. Steorts<sup>1</sup>

Duke University,<sup>1</sup> Bocconi University,<sup>2</sup> Microsoft Research,<sup>3</sup> and Harvard University<sup>4</sup>

## Summary

- Many models assume the number of data points in each cluster grows linearly with the total number of data points.
- Examples include infinitely exchangeable clustering processes.
- Entity resolution: the size of each cluster is often unrelated to the size of the data set.
- Requires models that yield clusters whose sizes grow sublinearly with the size of the data set.

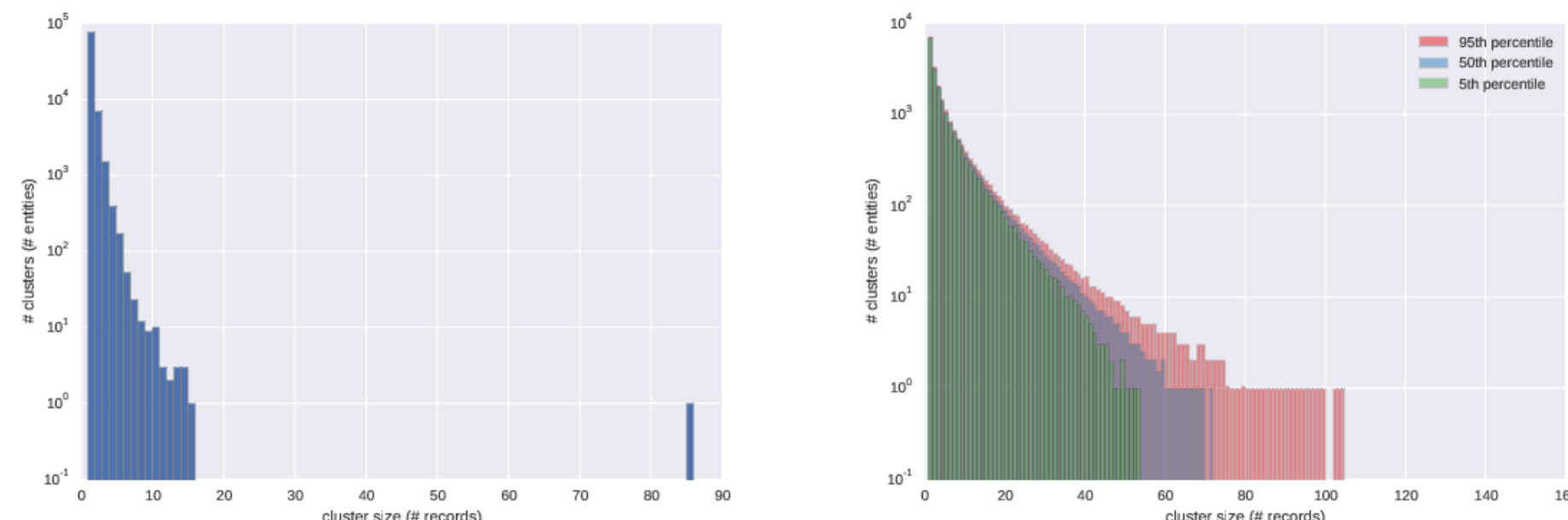


FIGURE 1: Two illustrations of the microclustering problem: (*Left*) a histogram of the number of records per entity (i.e., cluster sizes) for a database of 100,000 campaign finance donations from 2011-2012. (*Right*): A histogram of the cluster sizes generated from a Chinese restaurant process simulation with concentration parameter 0.1.

## Microclustering property

### Traditional Clustering

- A number of popular clustering applications assume priors on partitions such as the Dirichlet process and Pitman-Yor process distributions.
- Any* infinitely exchangeable partition distribution on the positive integers has a Kingman paintbox representation.
- Having a Kingman paintbox representation implies that the number of data points in each cluster grows (a.s.) linearly with the total number of data points  $N$ .
- By contrast, in entity resolution (and other problems), we expect the size of the clusters to be small even for large datasets.

### Microclustering

A sequence of random partitions exhibits the *microclustering property* if  $M_N$  is  $o_p(N)$ , where  $M_N \leq N$  is the size of the largest cluster in  $C_N$ , where  $C_N$  is a partition of  $[N] = \{1, 2, \dots, N\}$ .

- Equivalently,  $M_N / N \rightarrow 0$  in probability as  $N \rightarrow \infty$ .
- Microclustering requires sacrificing either (i) exchangeability or (ii) self-consistent marginalization of the partitions.
- Previous work Wallach et al. (2010) sacrificed (i); we sacrifice (ii).

## Flexible Models for Microclustering

### Notation

- $K$ : number of clusters (random).
- $N_k$ ,  $k = 1, \dots, K$ : size of the  $k$ th cluster.
- $N = \sum_{k=1}^K N_k$ : total number of data points.
- $z_n$ ,  $n = 1, \dots, N$ : cluster assignment of the  $n$ th data point.

Define

$$K \sim \kappa \quad \text{and} \quad N_1, \dots, N_K \mid K \stackrel{iid}{\sim} \mu. \quad (1)$$

where  $\kappa = (\kappa_1, \kappa_2, \dots)$  and  $\mu = (\mu_1, \mu_2, \dots)$  are probability distributions on  $\{1, 2, \dots\}$ . Given  $N_1, \dots, N_K$ , generate  $z_1, \dots, z_N$  by drawing a vector uniformly at random from the set of permutations of

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

### NBNB Model

Suppose that we allow

$$K \sim \text{Neg-Bin}(a, q) \quad \text{and} \quad N_1, \dots, N_K \mid K \sim \text{Neg-Bin}(r, p),$$

Note

- $a$  and  $q$  are assumed to be known.
- $r$  and  $p$  are sampled from Gamma and Beta priors.

We call the resulting marginal distribution of  $C_N$  the **NegBin-NegBin (NBNB)** model.

### NBD Model

For a fully nonparametric approach, allow

$$K \sim \text{Neg-Bin}(a, q) \quad \text{and} \quad N_1, \dots, N_K \mid K \sim \mu, \\ \mu \sim \text{Dirichlet}(\alpha, \mu^{(0)}),$$

with fixed concentration parameter  $\alpha > 0$  and known base measure

$$\mu^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \dots).$$

### Sampling

The reseating algorithm for the NBNB model is:

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

We can derive a similar reseating algorithm for the NBD model.

Compare to the reseating of the Chinese Restaurant Process (CRP):

- for  $n = 1, \dots, N$ , reassign element  $n$  to
  - an existing cluster  $c \in C_N \setminus n$  with probability  $\propto |c|$
  - a **new cluster** with probability  $\propto \alpha$ .

The parameter  $\alpha$  induces the richer get richer behavior in the CRP.

Main difference is the assignment to a new cluster depends on  $K$  through the term  $|C_N \setminus n|$ .

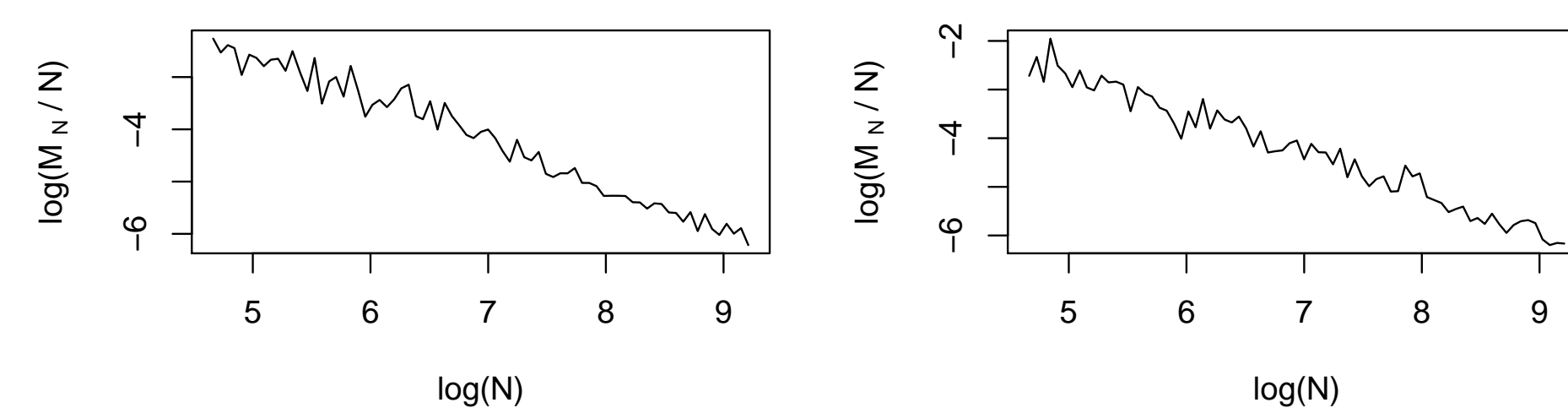


FIGURE 2: Empirical evidence suggesting that the NBNB (left) and NBD (right) models exhibit the microclustering property.

- We can use the reseating algorithms to draw samples from  $P(C_N \mid N)$  but they do not produce exact samples like the CRP.
- When the NBNB or NBD models are used as the prior in a partition-based clustering model, the resulting Gibbs sampling algorithm for  $C_N$  is similar to this algorithm accompanied by appropriate likelihood terms.
- Unfortunately, incremental Gibbs sampling is slow for large data sets.

## Inference

In real world clustering tasks, data points  $x_1, \dots, x_n$  and  $N$  are observed. Assume  $x_{n,\ell}$  where:

- $n = 1, \dots, N$  indexes how many records we observe.
- $\ell = 1, \dots, L$  indexes the categorical features within a record.

Let  $\zeta : \bigcup_{N=0}^{\infty} (C_N \times [N]) \rightarrow \{1, 2, \dots\}$  be a function that maps a partition  $C_N$  and a record  $x_{n,\ell}$  to its latent cluster assignment  $z_n$ .

$$C_N \sim \text{FMM}(\cdot) \\ z_n \mid C_N = \zeta(C_N, n) \\ \theta_{\ell,k} \sim \text{Dirichlet}(\delta_\ell, \gamma_\ell) \\ x_{n,\ell} \mid z_n, \theta_{\ell,1}, \theta_{\ell,2}, \dots \sim \text{Multinomial}(\theta_{\ell,z_n}),$$

- The base measure  $\gamma_\ell$  is assumed known.
- The concentration parameter  $\delta_\ell \sim \text{Gamma}(1, 1)$ .

## Chaperones Algorithm

If we let  $c_n \in C_N$  denote the cluster containing element  $n$ , then each iteration consists of:

- Randomly choose two *chaperones*,  $i, j \in \{1, \dots, N\}$  from a distribution  $P(i, j \mid x_1, \dots, x_N)$  where the probability of  $i$  and  $j$  given  $x_1, \dots, x_N$  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  $x_1, \dots, x_N$ .
- Reassign each  $n \in c_i \cup c_j$  by sampling from  $P(C_N \mid N, C_N \setminus n, c_i \cup c_j, x_1, \dots, x_N)$ .

Step 2 is almost identical to the restricted Gibbs moves found in existing split-merge algorithms (Jain and Neal, 2004), except that the chaperones  $i$  and  $j$  can also change clusters, provided they do not abandon any of their children.

## Experiments

### The Data

We assess how well each model “fits” on simulated data based on three real data sets (medical data, official statistics data, and the Syrian conflict).

- Italy Data: Survey data from Italy with 74% of data are singletons.
- NLTCS5000: Sample from the National Long Term Care Survey (NLTCS) with 68% singletons.
- Syria2000: Sample from Syrian conflict with 86% singletons.
- Ground truth is known for all the above data sets, so we compare using such unique identifiers.

### Evaluation Criterion

- Consider four statistics: the number of singleton clusters, the maximum cluster size, the mean cluster size, and the 90% quantile of cluster sizes.
- Consider three entity resolution metrics: estimated mean cluster size, False Negative Rate (FNR), and False Discovery Rate (FDR).
- Comparisons performed against the DP and PYP.

### Results

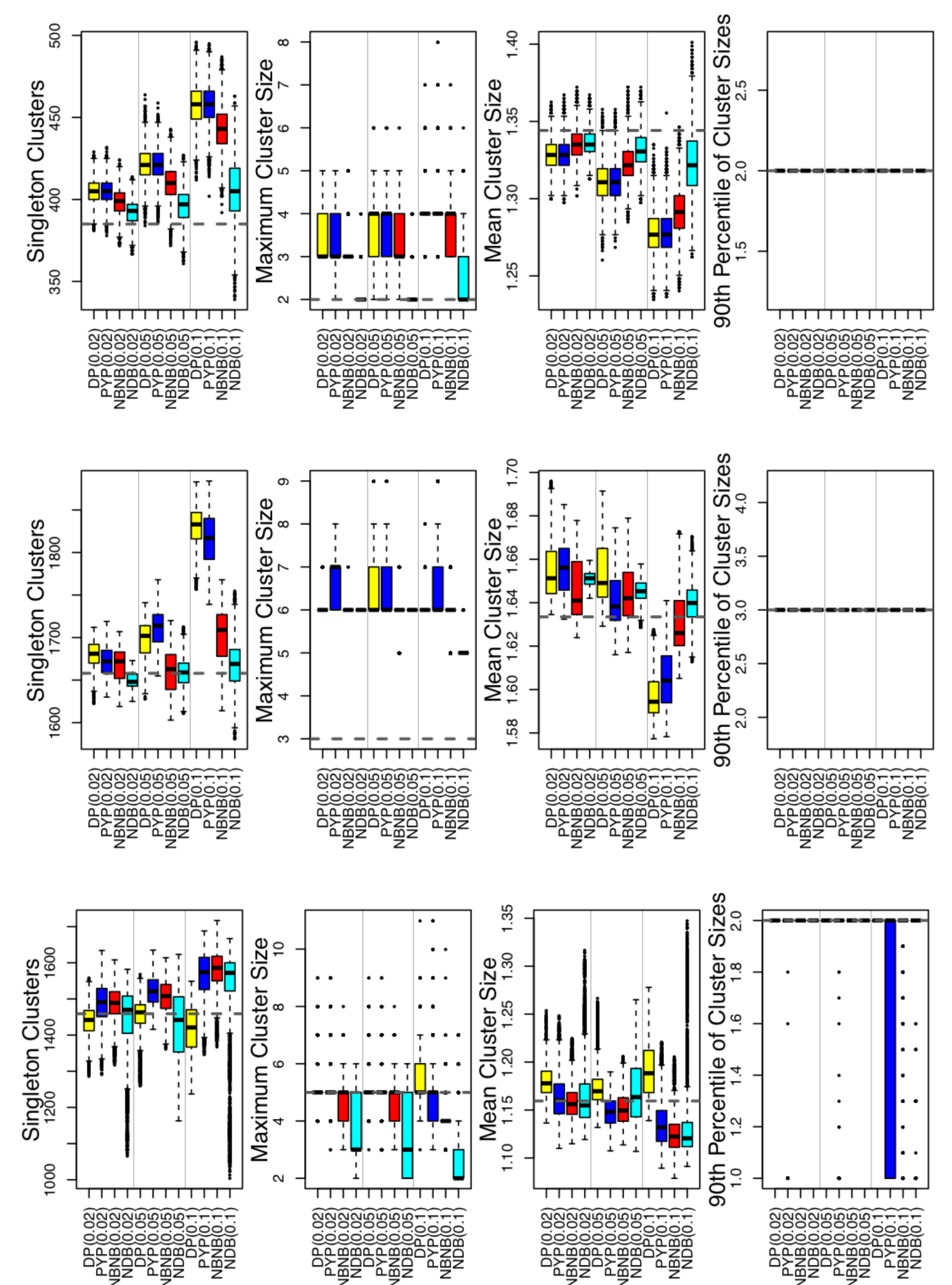


FIGURE 3: Top: Italy data. Middle: NLTCS5000. Bottom: Syria2000. The dashed horizontal line represents the true value of the statistic.

	$\hat{K}$	SD	FNR	FDR	$\hat{\delta}_\ell$
$DP_{\delta_\ell=0.02}$	594.00	4.51	0.07	0.03	0.02
$PY_{\delta_\ell=0.02}$	593.90	4.52	0.07	0.03	0.02
$NBNB_{\delta_\ell=0.02}$	591.00	4.43	0.04	0.03	0.02
$DNB_{\delta_\ell=0.02}$	590.50	3.64	0.03	0.00	0.02
$DP_{\delta_\ell=0.05}$	601.60	5.89	0.13	0.03	0.03
$PY_{\delta_\ell=0.05}$	601.50	5.90	0.13	0.03	0.04
$NBNB_{\delta_\ell=0.05}$	596.40	5.79	0.11	0.04	0.04
$DNB_{\delta_\ell=0.05}$	592.60	5.20	0.09	0.04	0.04

Table 1: Italy Data: Entity-resolution summary statistics and posterior expected value of  $\delta$ . The true number of clusters is  $K = 587$ .

## Discussion

- NBD and NBNB do as well or better than PYP and DP for real data sets.
- Expect the FMM models to beat the PYP and DP models as  $N$  increases.
- Ongoing work to string-valued features for inference is in progress.

**Acknowledgements:** This work was supported in part by NSF grants SBE-0965436, DMS-1045153, and IIS-1320219; NIH grant 5R01ES017436-05; the John Templeton Foundation; the Foerster-Bernstein Postdoctoral Fellowship; the UMass Amherst CIIR; and an EPSRC Doctoral Prize Fellowship.