Microclustering Models for Record Linkage Tasks

Brenda Betancourt

Foerster-Bernstein Postdoctoral Fellow
Department of Statistical Science
Duke University

joint work with Giacomo Zanella, Jeff Miller,
Hanna Wallach, Abbas Zaidi, Rebecca Steorts

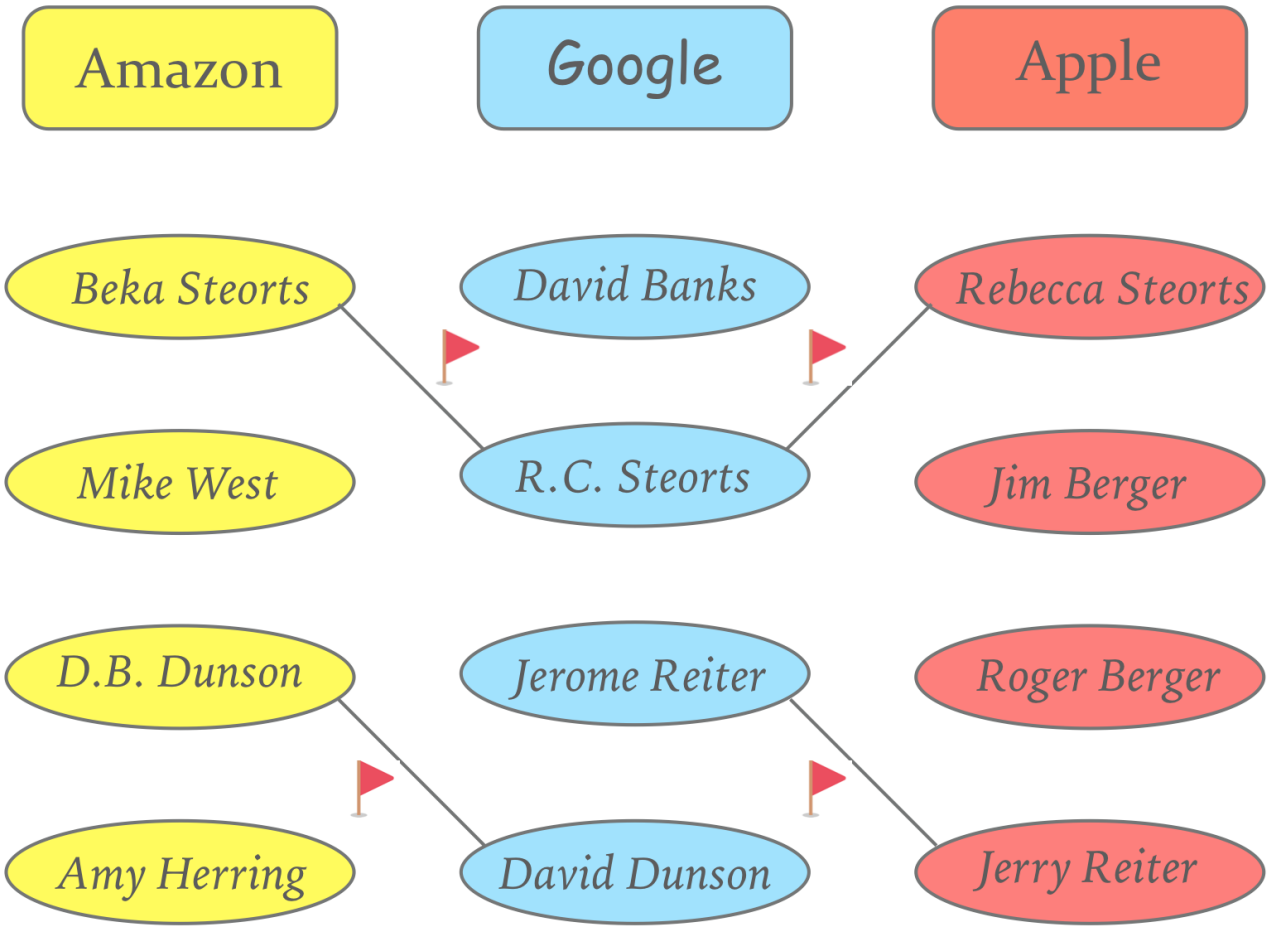
December 7, 2017

Record Linkage

- Record linkage: integration of multiple data sources removing duplicated information.
- Duplicated record: corrupted, degraded, or noisy representations of the same entity.
- Fields in the data can contain lies, omissions, typographical errors or out-of-date information.
- Particularly difficult when there are no unique identifiers across databases or anonymized data.

Record Linkage

Which records correspond to the same person?



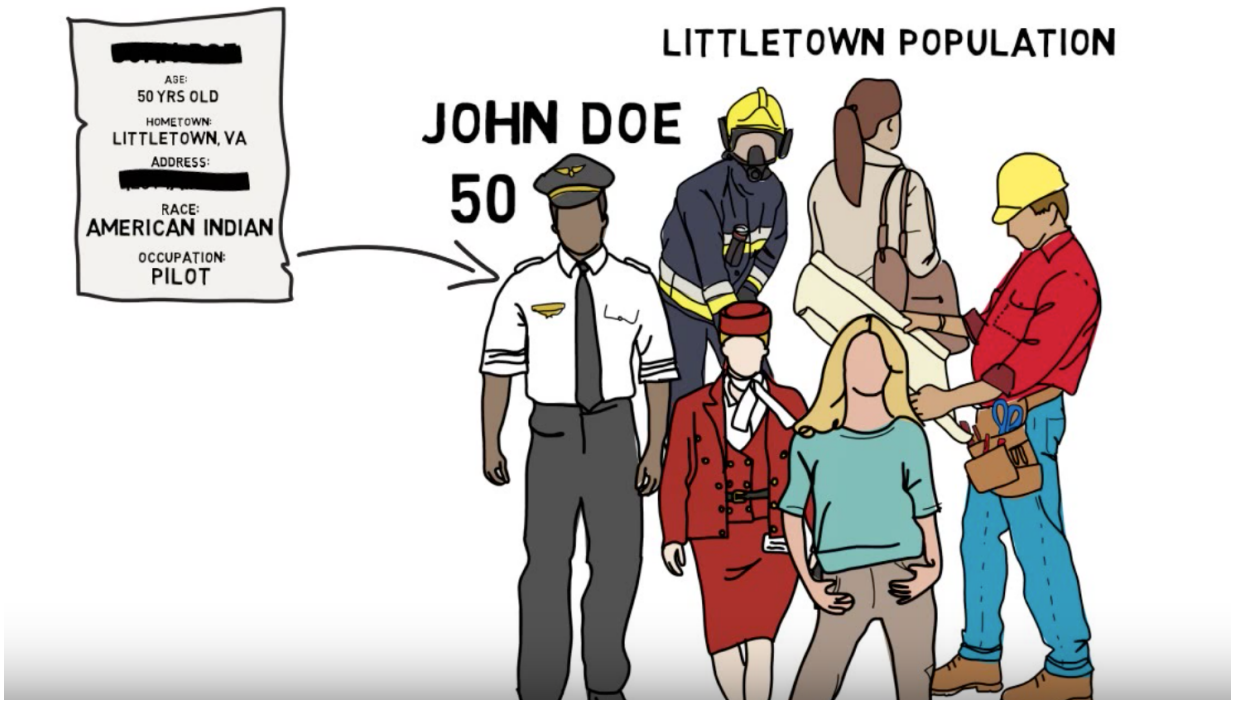
Motivation: Public Health

HIPAA
Health Insurance Portability and Accountability Act

- Protect people from losing their health insurance due to job change or pre-existing conditions.
- Reduce costs and administrative burdens by using electronic formats (EHR or EMR).
- Privacy of individually identifiable health information.

Data Confidentiality

- Limited access to unique identifiers e.g., Social Security Numbers, names, address.
- Balance between disclosure risk and data utility for record linkage purposes.



Record Linkage Feasibility

- Reliable and accurate linkage depends greatly on the quantity and quality of the identifying information that overlaps in the data sources being linked.
- Identifying records uniquely is very difficult when the proportion of duplicates is low with respect to the number of records and/or the discriminatory power of the identifiers is low (see Johndrow et al., 2017 and Steorts et al., 2017).
- The main focus here is estimation of the number of unique records in the data.

Probabilistic Record Linkage

- Fellegi and Sunter (1969): record pairs are a match, possible match, or non-match based on the calculation of linkage scores and likelihood ratio tests.
  - Only for two databases and no transitive closures:

    Pair records *(A, B)* and *(B,C)* are declared to be a match, but *A* and *C* are declared a non-match

A	Mary Smith	123 Oak	555-1234
B	Mary Smith	456 Elm	555-1234
C	Mary Smith	456 Elm	null

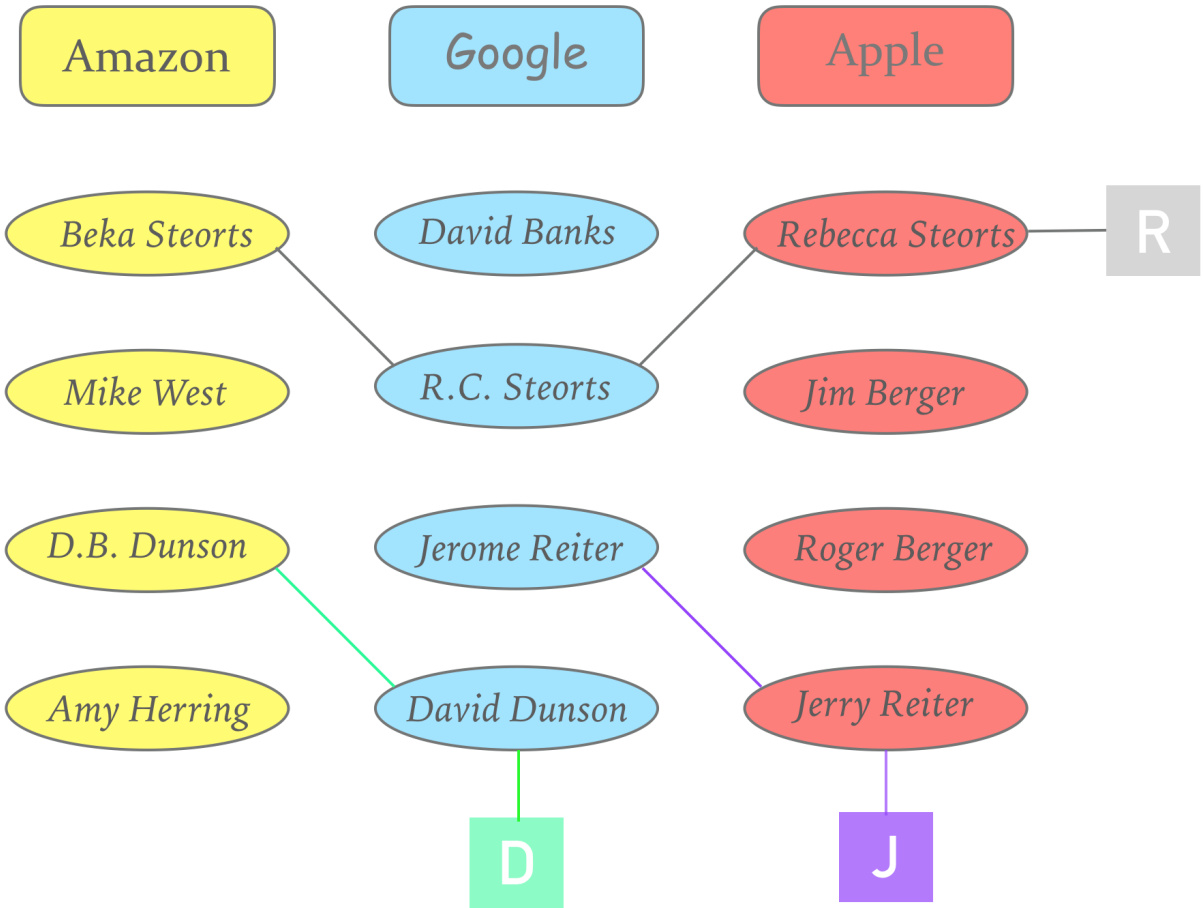
- Sadinle and Fienberg (2013) allow linkage of multiple data files and transitive closures.

Recent Bayesian methods

- Models that are reliable and account for the uncertainty of the record linkage process.
- Tancredi and Liseo (2011): discrete fields, both record linkage and population estimation in capture-recapture scenario.
- Gutman et. al (2013): record linkage as a missing data problem for medical data.
- Sadinle (2014), Steorts et al. (2014,2016): bipartite graph for latent entities creating a partition of the records.

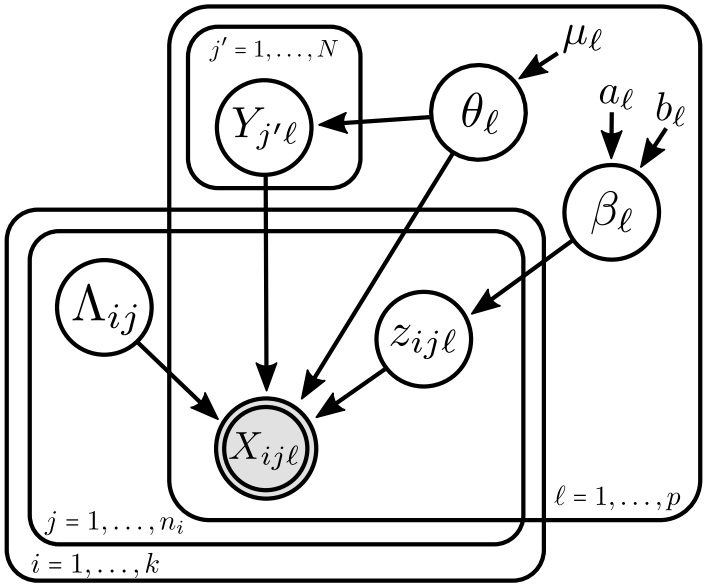
Record Linkage and Clustering

- 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):



- Λ_{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

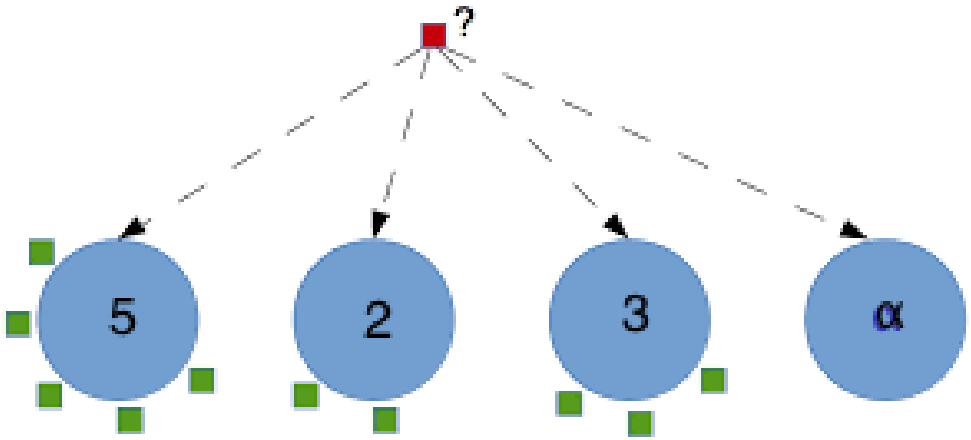
Goal: cluster N data points $x_1, \cdots, x_N$ into K clusters.

- Place a prior 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.

DP Mixture Models

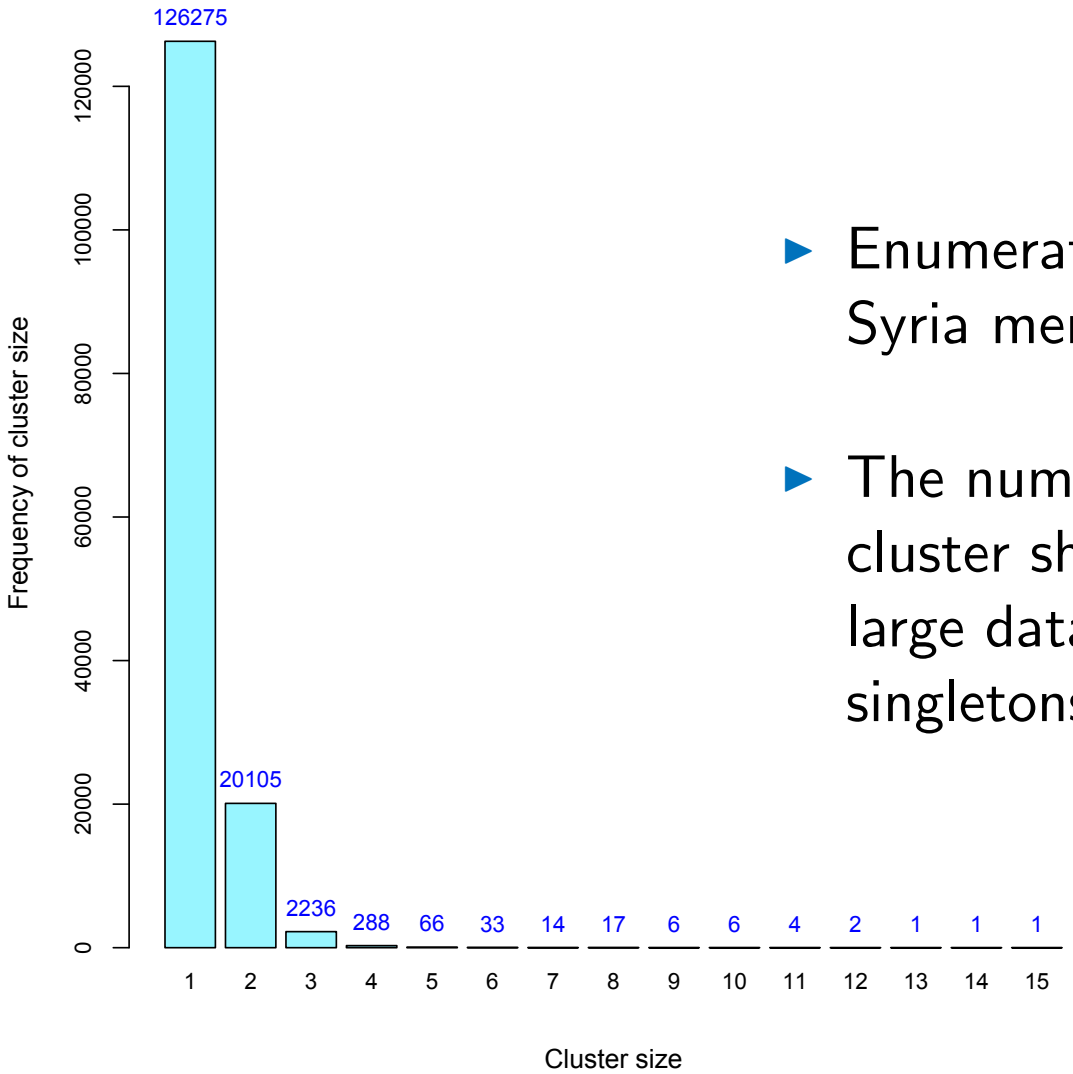
Other clustering tasks require models that assume cluster sizes grow linearly with the size of the data set.

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



- Pitman-Yor Process: Power-law tail

Why 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.

Microclustering property

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if $M_N/N \to 0$ in probability as $N \to \infty$, where M_N is the size of the largest cluster in C_N.

- As the number of data points grows, the size of each cluster is negligible.

Microclustering property

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if $M_N/N \to 0$ in probability as $N \to \infty$, where M_N is the size of the largest cluster in C_N.

- As the number of data points grows, the size of each cluster is negligible.
- To obtain the microclustering property, we must sacrifice either
  (a) C_N is exchangeable
  (b) The sequence is consistent in distribution
- We sacrifice (b)

Betancourt, B., Zanella, G., Wallach, H., Miller, J., Zaidi, A. and Steorts, R. (2016). Flexible Models for Microclustering with Applications to Entity Resolution, Advances in Neural Information Processing Systems (NIPS), Vol. 29, pp 1417-1425.

Kolchin Partition Models (KPM)

Define

$$K \sim \boldsymbol{\kappa} \quad \text{and} \quad N_1, \ldots, N_K \mid K \overset{iid}{\sim} \boldsymbol{\mu}.$$

where $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots)$ are probability distributions on $\{1, 2, \ldots\}$.

- N_k represents the size of cluster k i.e. $N = \sum_{k=1}^{K} N_k$

Kolchin Partition Models (KPM)

Define

$$K \sim \boldsymbol{\kappa} \quad \text{and} \quad N_1, \ldots, N_K \mid K \overset{iid}{\sim} \boldsymbol{\mu}.$$

where $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots)$ are probability distributions on $\{1, 2, \ldots\}$.

- N_k represents the size of cluster k i.e. $N = \sum_{k=1}^{K} N_k$
- Given the cluster sizes, the cluster assignments $z_1, \ldots, z_N$ are drawn uniformly at random from the set of permutations of

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

- Appropriate choices of $\boldsymbol{\kappa}$ and $\boldsymbol{\mu} \rightarrow$ Microclustering property

The NBNB Model

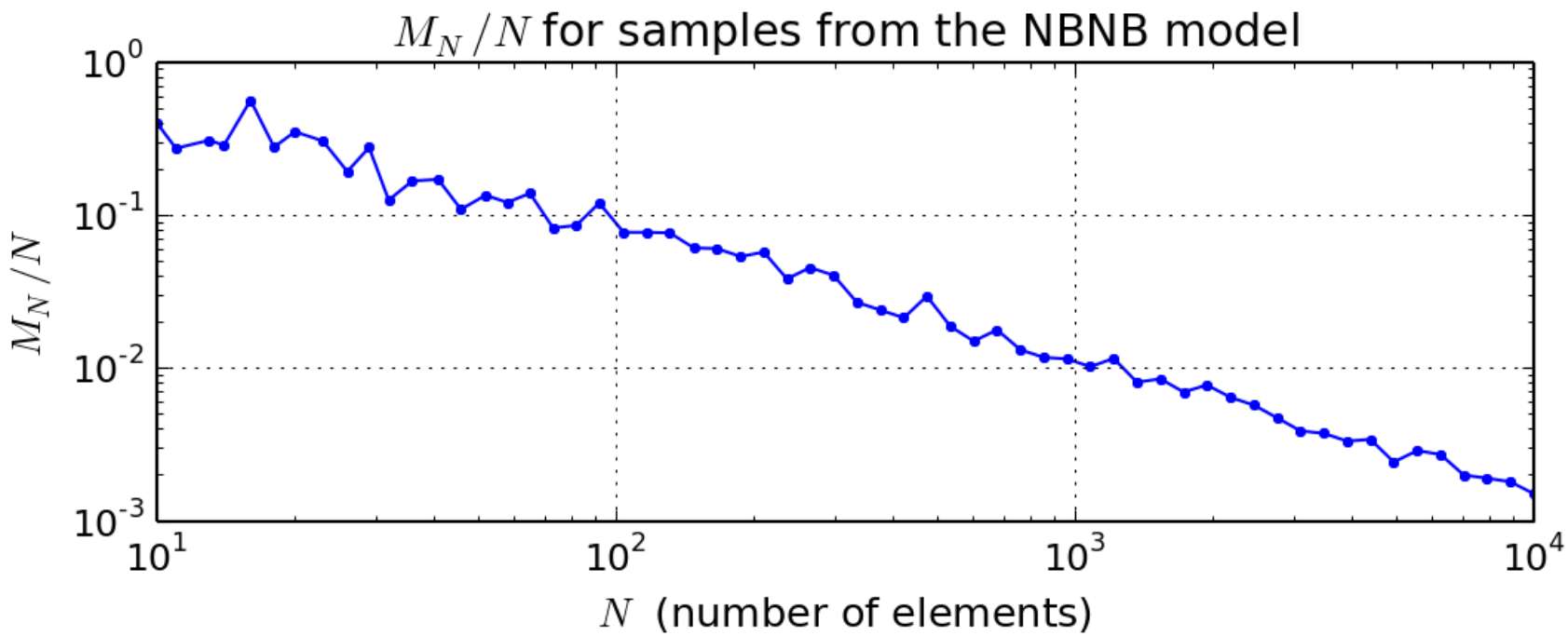
- We assume

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

- The reseating algorithm for the NBNB model is:
  - For $n = 1, \ldots N$, reassign element n to
    - an existing cluster $c \in C_N \setminus n$ with probability $\propto |c| + r$
    - a new cluster with probability $\propto (|C_N \setminus n| + a) r\beta$.
- Sampling is similar to the Dirichlet Process except no exact samples.

NBNB Model: Microclustering property

- $a = 1,\ q = 0.9$
- r and p such that $E(N_k|K) = 3$ $var(N_k|K) = 3^2/2$



The NBD Model

We change the model formulation to

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

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

$$\boldsymbol{\mu}^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \cdots),$$

with $\sum_{m=1}^{\infty} \mu_m^{(0)} = 1$ and $\mu_m^{(0)} \geq 0$.

Record Linkage with Categorical Data

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 $x_{n,\ell}$ to its latent cluster assignment z_n.

$$
\begin{aligned}
C_N &\sim \text{KPM}(\cdot) \\
\boldsymbol{\theta}_{\ell,k} | \delta_\ell, \boldsymbol{\gamma}_\ell &\sim \text{Dir}(\delta_\ell, \boldsymbol{\gamma}_\ell) \\
z_n \mid C_N &= \zeta(C_N, n) \\
x_{n,\ell} \mid z_n, \boldsymbol{\theta}_{\ell,1}, \boldsymbol{\theta}_{\ell,2}, \ldots &\sim \text{Cat}(\boldsymbol{\theta}_{\ell,z_n}),
\end{aligned}
$$

- The base measure $\boldsymbol{\gamma}_\ell$ is assumed known.
- Assume a Gamma prior for the concentration parameter δ_ℓ.

Experiments: Data

1. Italy Data: Italian Survey on Household Income and Wealth (SHIW) from Friuli region.
   - N=789, 74% singleton clusters, nine fields (year of birth, sex, working status, employment status, branch of activity, town size, geographical area of birth, whether or not Italian national, and highest educational level)

2. NLTCS5000: National Long Term Care Survey (NLTCS).
   - N=5000, 68% singleton clusters, seven fields (date of birth, sex, state, regional office)

3. SyriaSizes: Data collected from 4 human rights organizations on killings in the Syrian war between 2011 and 2014.
   - N=6700, 61% singleton clusters, five fields (date of dead, sex, governorate)

Experiments: Setup

- a and q are assumed to be known.
- r and p are sampled from Gamma(1,1) and Beta(2,2) priors.
- Choose α and $\boldsymbol{\mu}^{(0)}$ (e.g. $\alpha = 1$, $\boldsymbol{\mu}^{(0)} \sim$ Geom(p))

Experiments: Setup

- a and q are assumed to be known.
- r and p are sampled from Gamma(1,1) and Beta(2,2) priors.
- Choose α and $\boldsymbol{\mu}^{(0)}$ (e.g. $\alpha = 1$, $\boldsymbol{\mu}^{(0)} \sim$ Geom(p))

Performance evaluation:

- Comparison with Dirichlet Process and Pitman-Yor Process.
- Classification error: False Negative Rate (FNR) and False Discovery Rate (FDR)

$$FNR = \frac{FN}{CL + FN} \qquad FDR = \frac{FP}{CL + FP}$$

FN: False Negatives; FP: False Positives; CL: Correct Links

Inference on Italy Data

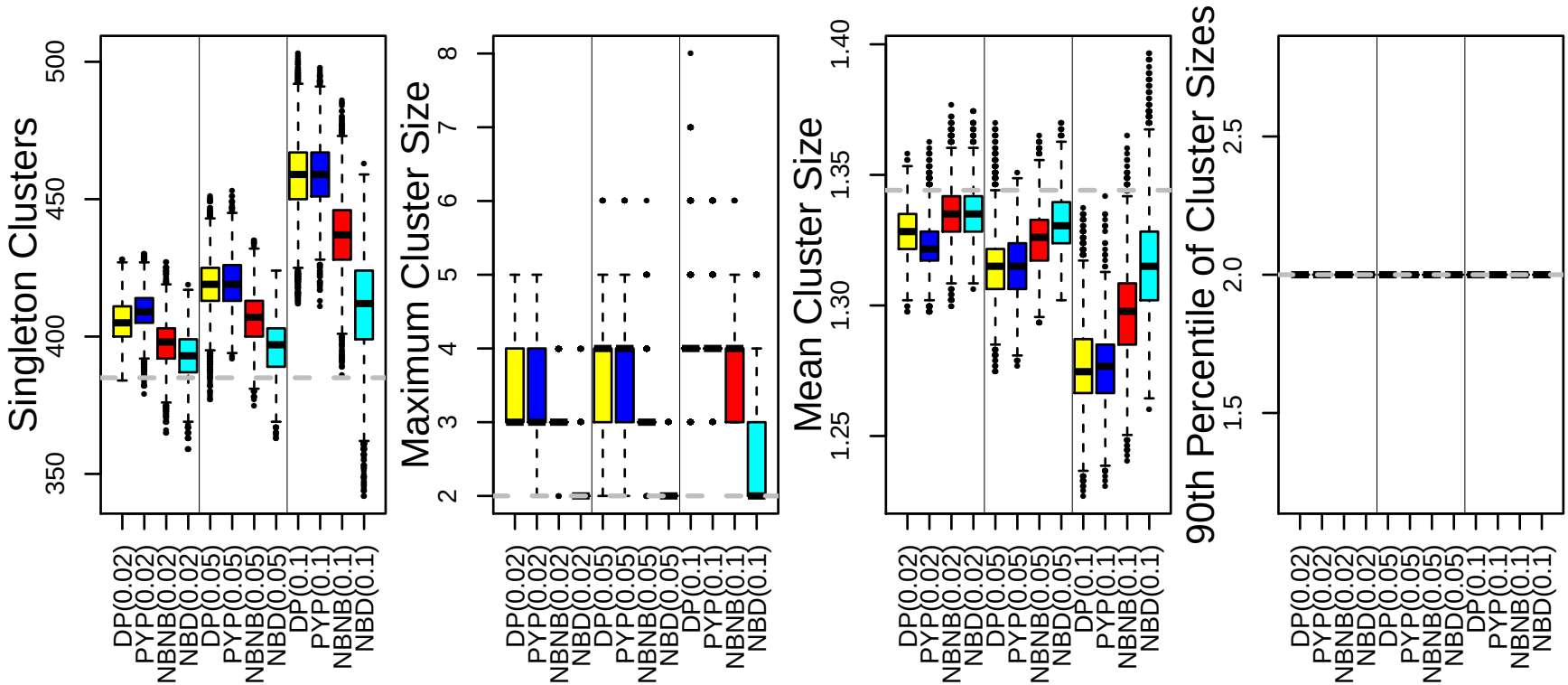


Figure: NBD > NBNB > PYP > DP

Inference on Italy Data: True K=587

	$\hat{K}$	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
$NBD_{\delta_\ell=0.02}$	590.50	3.64	0.03	0.00	0.02

Inference on Italy Data: True K=587

	$\hat{K}$	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
$NBD_{\delta_\ell=0.02}$	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
$NBD_{\delta_\ell=0.05}$	592.60	5.20	0.09	0.04	0.04

Inference on Italy Data: True K=587

	$\hat{K}$	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
$NBD_{\delta_\ell=0.02}$	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
$NBD_{\delta_\ell=0.05}$	592.60	5.20	0.09	0.04	0.04
$DP_{\delta_\ell=0.10}$	617.40	7.23	0.27	0.06	0.07
$PY_{\delta_\ell=0.10}$	617.40	7.22	0.27	0.05	0.07
$NBNB_{\delta_\ell=0.10}$	610.90	7.81	0.24	0.06	0.08
$NBD_{\delta_\ell=0.10}$	596.60	9.37	0.18	0.05	0.10

Inference on NLTCS5000

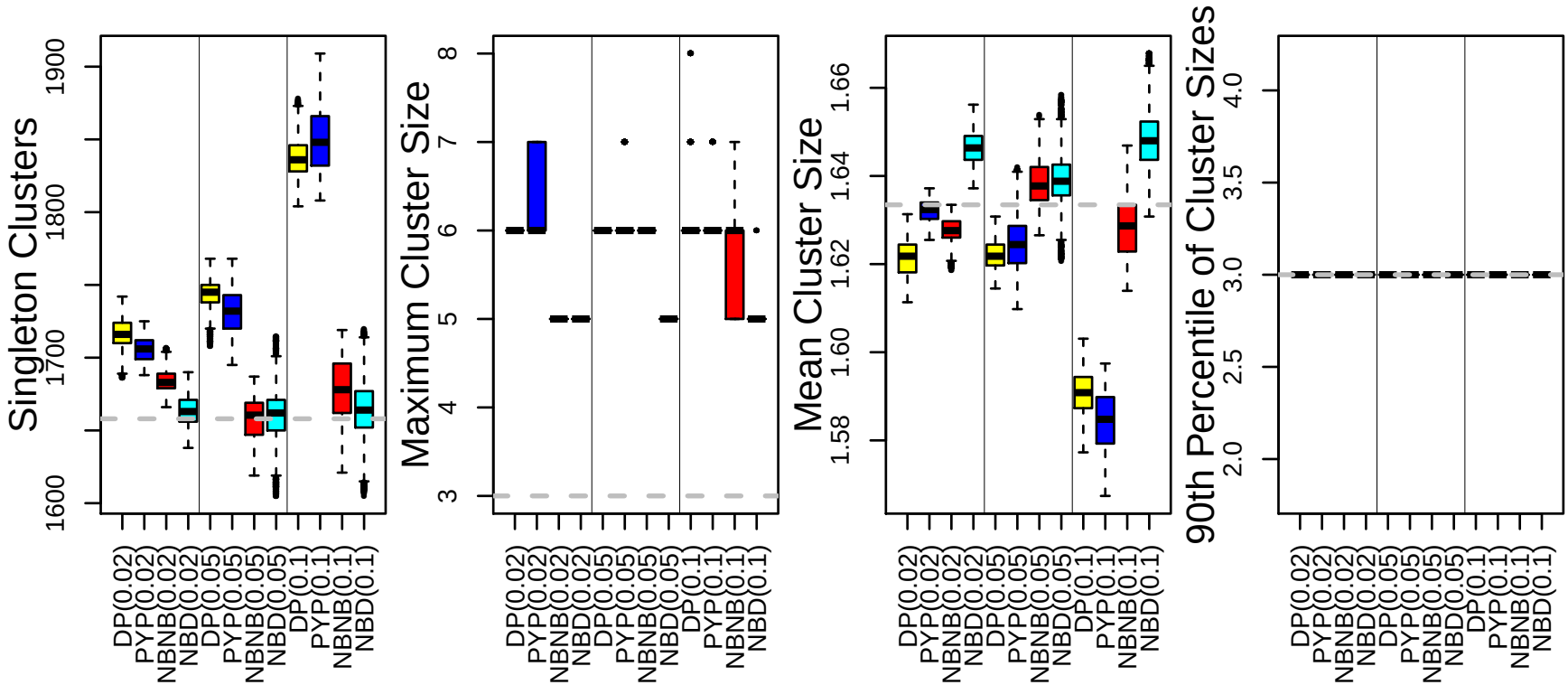


Figure: NBD $\geq$ NBNB $>$ PYP $>$ DP

Inference on NLTCS5000: True K=3061

	$\hat{K}$	SD	FNR	FDR	$\hat{\delta}_\ell$
$DP_{\delta_\ell=0.02}$	3021.70	24.96	0.02	0.11	0.03
$PY_{\delta_\ell=0.02}$	3018.70	25.69	0.03	0.11	0.03
$NBNB_{\delta_\ell=0.02}$	3037.80	25.18	0.02	0.07	0.02
$NBD_{\delta_\ell=0.02}$	3028.20	5.65	0.01	0.09	0.03
$DP_{\delta_\ell=0.10}$	3130.50	21.44	0.12	0.09	0.10
$PY_{\delta_\ell=0.10}$	3115.10	25.73	0.13	0.10	0.10
$NBNB_{\delta_\ell=0.10}$	3067.30	25.31	0.11	0.08	0.11
$NBD_{\delta_\ell=0.10}$	3049.10	16.48	0.09	0.08	0.12

Inference on SyriaSizes

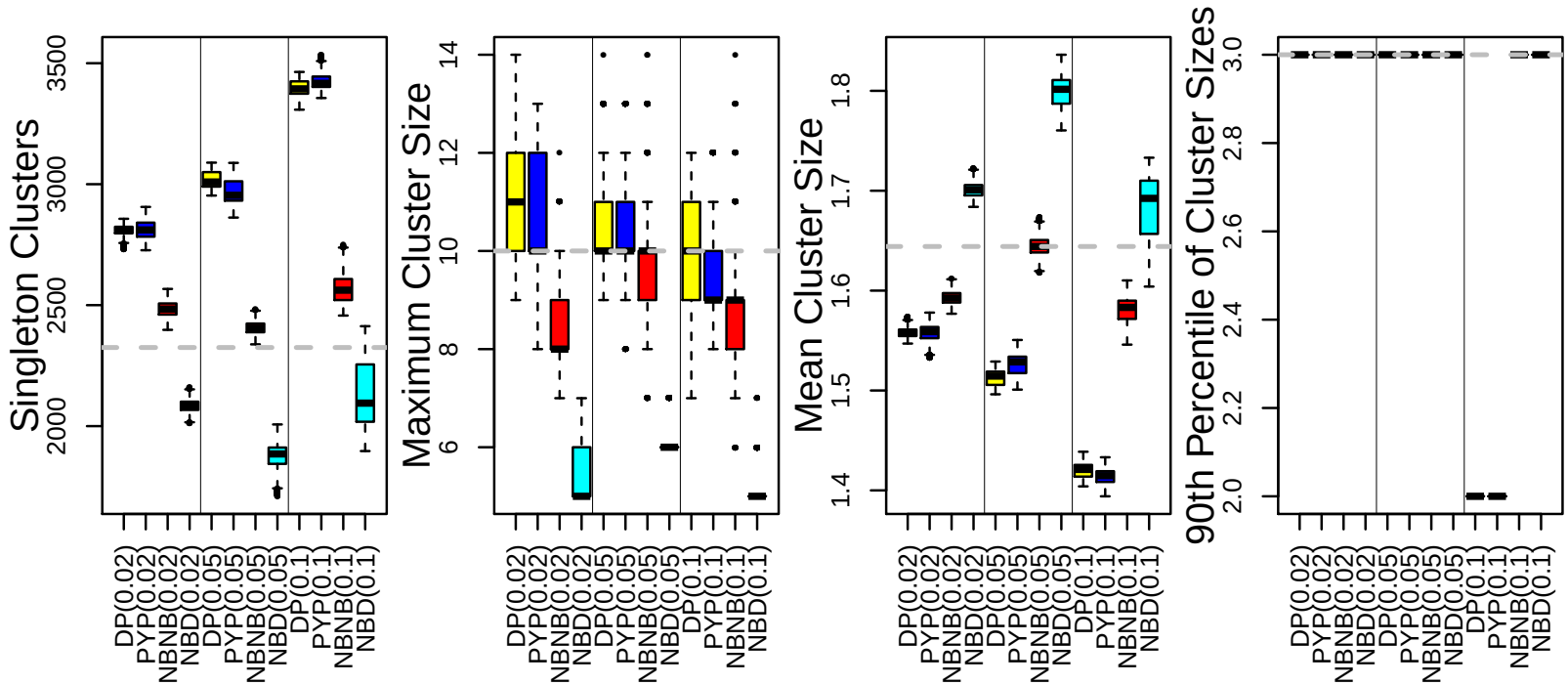


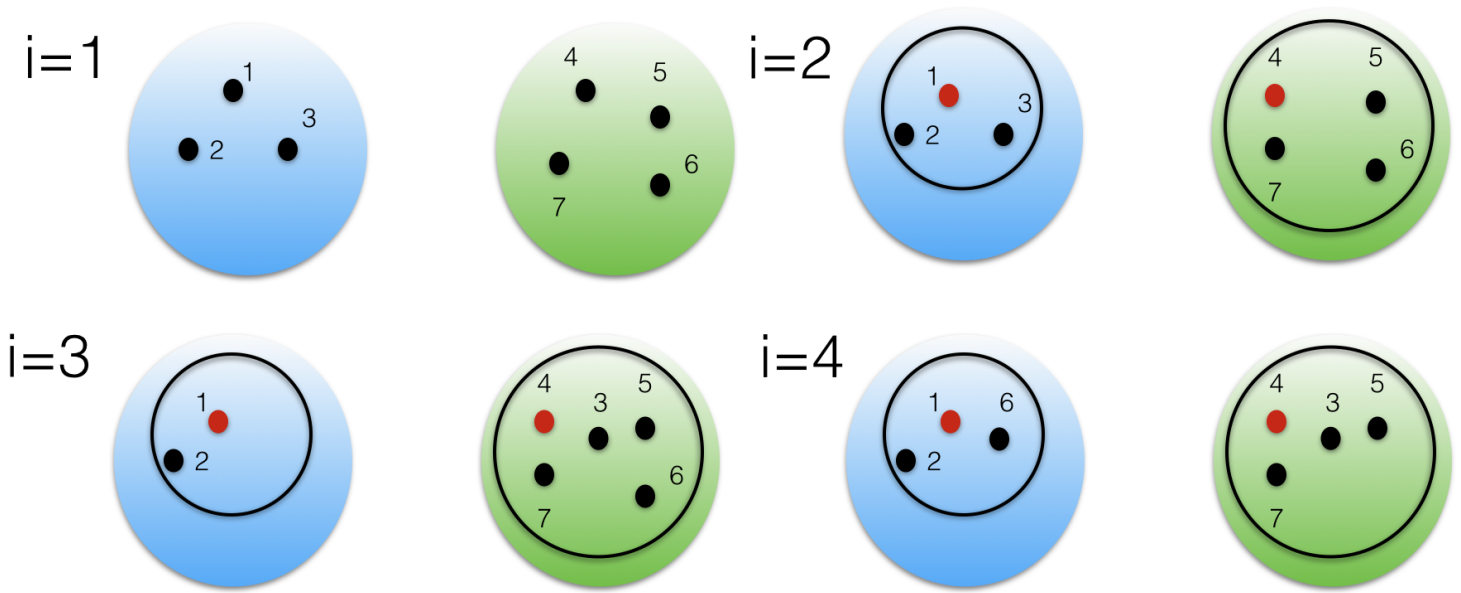
Figure: No clear winner here.

Inference on SyriaSizes: True K=4,075

	$\hat{K}$	SD	FNR	FDR	$\hat{\delta}_\ell$
$DP_{\delta_\ell=0.02}$	4175.70	66.04	0.65	0.17	0.01
$PY_{\delta_\ell=0.02}$	4234.30	68.55	0.64	0.19	0.01
$NBNB_{\delta_\ell=0.02}$	4108.70	70.56	0.65	0.19	0.01
$NBD_{\delta_\ell=0.02}$	3979.50	70.85	0.68	0.20	0.03
$DP_{\delta_\ell=0.10}$	4507.40	82.27	0.80	0.19	0.03
$PY_{\delta_\ell=0.10}$	4540.30	100.53	0.80	0.20	0.03
$NBNB_{\delta_\ell=0.10}$	4400.60	111.91	0.80	0.23	0.03
$NBD_{\delta_\ell=0.10}$	4251.90	203.23	0.82	0.25	0.04

Computation

- Incremental Gibbs is very slow → Chaperones algorithm: similar to the split-merge MCMC (Jain and Neal, 2004)



Miller, J.W., **Betancourt**, B., Zaidi, A., Walach, H., Steorts, R. (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*.

Traditional Blocking

Blocking reduces the comparison space to record pairs that meet certain basic criteria.

- Match only pairs that agree on e.g. sex and month of birth.
- Record pairs that do not meet the blocking criteria are automatically classified as non-matches.
- Fields can be unreliable for many applications and blocking may miss large proportions of matches i.e. increased false negatives rates.
- Trade-off between block sizes: true matches being missed vs computational efficiency.

Conclusions and Ongoing Work

- We introduce a new class of prior models to address the microclustering problem common to record linkage tasks.
- Theoretical properties to fully characterize the distribution of cluster sizes and frequencies for the microclustering models.
- Study performance bounds for record linkage feasibility with microclustering models.
- Scalability is still an issue for big data $\rightarrow$ better proposal distribution for the chaperones algorithm and blocking.
- R package for microclustering coming soon.

Other Research

- ## Record Linkage

  - Zanella, G., **Betancourt, B.**, Steorts, R.C. (2018). Exchangeable sequences of finite clusters. In preparation.

- ## Network Analysis

  - **Betancourt, B.**, Rodríguez, A., Boyd, N. (2017+). Modeling and Prediction of Financial Trading Networks: A case study in the NYMEX natural gas futures market. In revision for Journal of Royal Statistical Society, Series C.

  - **Betancourt, B.**, Rodríguez, A., Boyd, N. (2017). Investigating Competition in Financial Markets: A sparse autologistic model for dynamic network data. Journal of Applied Statistics. In Press.

  - **Betancourt, B**., Rodríguez, A., Boyd, N. (2017). Bayesian Fused Lasso regression for dynamic binary networks. Journal of Computational and Graphical Statistics. In Press.

- ## Robust Priors

  - Fúquene, J., **Betancourt, B.**, Pereira, J. B. M. (2017). A weakly informative prior for Bayesian dynamic model selection with applications in fMRI. Journal of Applied Statistics. In Press.

References

Betancourt, B., Zanella, G., Wallach, H., Miller, J., Zaidi, A. and Steorts, R. (2016). Flexible Models for Microclustering with Applications to Entity Resolution, Advances in Neural Information Processing Systems (NIPS), Vol. 29, pp 1417-1425.

Miller, J.W., **Betancourt**, B., Zaidi, A., Walach, H., Steorts, R. (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*.

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).

Steorts R.C., Barnes M. and Neiswanger W.. Performance Bounds for Graphical Record Linkage. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:298-306, 2017.

Johndrow J.E., Lum K., Dunson D.B. (2017). Theoretical Limits of Record Linkage and Microclustering. arXiv:1703.04955v1.

Fellegi, I. and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64 1183-1210.

Gutman, R., Afendulis, C. and Zaslavsky, A. (2013). A bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108 34-47.

Tancredi, A. and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5 1553-1585.

Sadinle, M. and Fienberg, S. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems. Journal of the American Statistical Association, 108 385?397.

Questions? bb222@stat.duke.edu

Thank you to Patrick Ball and Megan Price at the Human Rights Data Analysis Group for providing the Syrian conflict data.

Thank you to the Foerster-Bernstein Postdoctoral Fellowship, the National Science Foundation for NSF CAREER and NSF Big Data Privacy grants, and the Laboratory for Analytic Sciences at NC State University for their support. The views in this talk are of the authors alone and not of the funding organizations.

Backup Slides

Microclustering and Mixture Models

- Kingman's paintbox theorem implies that any exchangeable partition of $\mathbb{N}$ is

  (a) either equal to the trivial partition in which each part contains one element or

  (b) satisfies $\liminf_{N\to\infty} M_N / N > 0$ with positive probability.

Microclustering and Mixture Models

- 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]

- 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).

The NBNB Model

Suppose that we allow

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

The marginal distribution of C_N is:

$$P(C_N|N, a, q, r, p) \propto \Gamma(C_N + a)\beta^{|C_N|} \prod_{c \in C_N} \frac{\Gamma(|c| + r)}{\Gamma(r)}$$

where $\beta = \frac{q(1-p)^r}{1-(1-p)^r}$.

The NBD Model

The marginal distribution of C_N is:

$$P(C_N|N, a, q, \boldsymbol{\mu}) \propto \Gamma(C_N + a)q^{|C_N|} \prod_{c \in C_N} |c|!\mu_{|c|}$$

We can derive a similar reseating algorithm :

- for $n = 1, \ldots N$, reassign element n to
  - an existing cluster $c \in C_N \setminus n$ with probability $\propto (|c| + 1)\frac{\mu_{(|c|+1)}}{\mu_{|c|}}$
  - a new cluster with probability $\propto (|C_N \setminus n| + a)q\mu_1$.

Note

- a and q are assumed to be known.
- Choose α and $\boldsymbol{\mu}^{(0)}$ (e.g. $\alpha = 1$, $\boldsymbol{\mu}^{(0)} \sim$ Geom(p))