Flexible Models for Microclustering with Application to Entity Resolution

Authored by:

Jeffrey W. Miller
Brenda Betancourt
Giacomo Zanella
Hanna Wallach
Abbas Zaidi
Rebecca C. Steorts

Abstract

Most generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman–Yor process mixture models make this assumption, as do all other infinitely exchangeable clustering models. However, for some applications, this assumption is inappropriate. For example, when performing entity resolution, the size of each cluster should be unrelated to the size of the data set, and each cluster should contain a negligible fraction of the total number of data points. These applications require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the microclustering property and introducing a new class of models that can exhibit this property. We compare models within this class to two commonly used clustering models using four entity-resolution data sets.

1 Paper Body

Many clustering applications require models that assume cluster sizes grow linearly with the size of the data set. These applications include topic modeling, inferring population structure, and discriminating among cancer subtypes. Infinitely exchangeable clustering models, including finite mixture models, Dirichlet process mixture models, and Pitman?Yor process mixture models, all make this lineargrowth assumption, and have seen numerous successes when used in these contexts. For other clustering applications, such as entity resolution, this assumption is inappropriate. Entity resolution (including record linkage and de-duplication) involves identifying duplicate2 records in noisy databases [1, 2],

traditionally by directly linking records to one another. Unfortunately, this traditional approach is computationally infeasible for large data sets?a serious limitation in ?the age of big data? [1, 3]. As a ?

Giacomo Zanella and Brenda Betancourt are joint first authors. In the entity resolution literature, the term ?duplicate records? does not mean that the records are identical, but rather that the records are corrupted, degraded, or otherwise noisy representations of the same entity. 2

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

result, researchers increasingly treat entity resolution as a clustering problem, where each entity is implicitly associated with one or more records and the inference goal is to recover the latent entities (clusters) that correspond to the observed records (data points) [4, 5, 6]. In contrast to other clustering applications, the number of data points in each cluster should remain small, even for large data sets. Applications like this require models that yield clusters whose sizes grow sublinearly with the total number of data points [7]. To address this requirement, we define the microclustering property in section 2 and, in section 3, introduce a new class of models that can exhibit this property. In section 4, we compare two models within this class to two commonly used infinitely exchangeable clustering models.

2

The Microclustering Property

To cluster N data points x1 , . . . , xN using a partition-based Bayesian clustering model, one first places a prior over partitions of [N] = {1, . . . , N }. Then, given a partition CN of [N], one models the data points in each part c ? CN as jointly distributed according to some chosen distribution. Finally, one computes the posterior distribution over partitions and, e.g., uses it to identify probable partitions of [N]. Mixture models are a well-known type of partition-based Bayesian clustering model, in which CN is implicitly represented by a set of cluster assignments z1 , . . . , zN . These cluster assignments can be regarded as the first N elements of an infinite sequence z1 , z2 , . . ., drawn a priori from ??H

iid

and z1 , z2 , . . . — ? ? ?,

(1)

P where H is a prior over ? and ? is a vector of mixture weights with l ?l = 1 and ?l ? 0 for all l. Commonly used mixture models include (a) finite mixtures where the dimensionality of ? is fixed and H is usually a Dirichlet distribution; (b) finite mixtures where the dimensionality of ? is a random variable [8, 9]; (c) Dirichlet process (DP) mixtures where the dimensionality of ? is infinite [10]; and (d) Pitman?Yor process (PYP) mixtures, which generalize DP mixtures [11]. Equation 1 implicitly defines a prior over partitions of N = {1, 2, . . .}. Any random partition CN of N induces a sequence of random partitions (CN : N = 1, 2, . . .), where CN is a partition of [N]. Via the strong law of large numbers, the cluster sizes in any such sequence obtained via equation 1 grow PN linearly with N because, with probability one, for all l, N1 n=1 I(zn = l)

? ?l as N ? ?, where I(?) denotes the indicator function. Unfortunately, this linear growth assumption is not appropriate for entity resolution and other applications that require clusters whose sizes grow sublinearly with N . To address this requirement, we therefore define the microclustering property: A sequence of random partitions (CN : N = 1, 2, . . .) exhibits the microclustering property if MN is op (N), where MN is the size of the largest cluster in CN , or, equivalently, if MN / N ? 0 in probability as N ? ?. A clustering model exhibits the microclustering property if the sequence of random partitions implied by that model satisfies the above definition. No mixture model can exhibit the microclustering property (unless its parameters are allowed to vary with N). In fact, Kingman?s paintbox theorem [12, 13] implies that any exchangeable partition of N, such as a partition obtained using equation 1, is either equal to the trivial partition in which each part contains one element or satisfies lim infN ?? MN / N ¿ 0 with positive probability. By Kolmogorov?s extension theorem, a sequence of random partitions (CN : N = 1, 2, . . .) corresponds to an exchangeable random partition of N whenever (a) each CN is finitely exchangeable (i.e., its probability is invariant under permutations of {1, . . . , N }) and (b) the sequence is projective (also known as consistent in distribution)?i.e., if N 0 ¡ N , the distribution over CN 0 coincides with the marginal distribution over partitions of [N 0] induced by the distribution over CN . Therefore, to obtain a nontrivial model that exhibits the microclustering property, we must sacrifice either (a) or (b). Previous work [14] sacrificed (a); in this paper, we instead sacrifice (b). Sacrificing finite exchangeability and sacrificing projectivity have very different consequences. If a partition-based Bayesian clustering model is not finitely exchangeable, then inference will depend on the order of the data points. For most applications, this consequence is undesirable?there is no reason to believe that the order of the data points is meaningful. In contrast, if a model lacks projectivity, then the implied joint distribution over a subset of the data points in a data set will not be the same as the joint distribution obtained by modeling the subset directly. In the context of entity resolution, sacrificing projectivity is a more natural and less restrictive choice than sacrificing finite exchangeability. 2

3

Kolchin Partition Models for Microclustering

We introduce a new class of Bayesian models for microclustering by placing a prior on the number of clusters K and, given K, modeling the cluster sizes N1 , . . . , NK directly. We start by defining K ? ? and

iid

N1 , . . . , NK — K ? ?,

(2)

where ? = (?1 , ?2 , . . .) and ? = (?1 , ?2 , . . .) are probability distributions over N = {1, 2, . . .}. PK We then define N = k=1 Nk and, given N1 , . . . , NK , generate a set of cluster assignments z1 , . . . , zN by drawing a vector uniformly at random from the set of permutations of (1, . . . , 1, 2, . . . , 2, ... , K, . . . , K). The cluster assignments z1 , . . . , zN induce a random par— {z } — {z } — {z } N1 times

N2 times

NK times

tition CN of [N], where N is itself a random variable?i.e., CN is a random partition of a random number of elements. We refer to the resulting class of marginal distributions over CN as Kolchin partition (KP) models [15, 16] because the form of equation 2 is closely related to Kolchin?s representation theorem for Gibbs-type partitions (see, e.g., 16, theorem 1.2). For appropriate choices of ? and ?, KP models can exhibit the microclustering property (see appendix B for an example). S? If CN denotes the set of all possible partitions of [N], then N =1 CN is the S?set of all possible partitions of [N] for all N ? N. The probability of any given partition CN ? N =1 CN is ! Y —CN —! ?—CN — P (CN) = —c—! ?—c— , (3) N! c?CN

where — ? — denotes the cardinality of a set, —CN — is the number of clusters in CN , and —c— is the number of elements in cluster c. In practice, however, N Conditioned on

Q is usually observed. N , a KP model implies that P (CN — N) ? —CN —! ?—CN — —c—! ? . Equation 3 leads to a —c— c?CN ?reseating algorithm??much like the Chinese restaurant process (CRP)?derived by sampling from P (CN — N, CN n), where CN n is the partition obtained by removing element n from CN : ? for n = 1, . . . , N , reassign element n to ? an existing cluster c ? CN n with probability ? (—c— + 1) ? or a new cluster with probability ? (—CN n— + 1)

?(—c—+1) ?—c—

?(—CNn—+1) ?—CNn— ?1 .

We can use this reseating algorithm to draw samples from P (CN — N); however, unlike the CRP, it does not produce an exact sample if it is used to incrementally construct a partition from the empty set. In practice, this limitation does not lead to any negative consequences because standard posterior inference sampling methods do not rely on this property. When a KP model is used as the prior in a partition-based clustering model?e.g., as an alternative to equation 1?the resulting Gibbs sampling algorithm for CN is similar to this reseating algorithm, but accompanied by likelihood terms. Unfortunately, this algorithm is slow for large data sets. In appendix C, we therefore propose a faster Gibbs sampling algorithm?the chaperones algorithm?that is particularly well suited to microclustering. In sections 3.1 and 3.2, we introduce two related KP models for microclustering, and in section 3.4 we explain how KP models can be applied in the context of entity resolution with categorical data. 3.1

The NBNB Model

We start with equation 3 and define ? = NegBin (a, q)

and

? = NegBin (r, p) ,

(4)

where NegBin(a, q) and NegBin(r, p) are negative binomial distributions truncated to N = {1, 2, . . . }. We assume that a ¿ 0 and q ? (0, 1) are fixed hyperparameters, while r and p are distributed as r ? Gam(?r , sr) and

p ? Beta(up , vp) for fixed ?r , sr , up and vp .3 We refer to the resulting marginal distribution over CN as the negative binomial?negative binomial (NBNB) model. 3

We use the shape-and-rate parameterization of the gamma distribution.

3

7

8

9

?4

?2 6

?6

log(M N / N)

?4 ?6

log(M N / N)

5

5

6

log(N)

7

8

9

log(N)

Figure 1: The NBNB (left) and NBD (right) models appear to exhibit the microclustering property. By substituting equation 4 into equation 3, we obtain the probability of CN conditioned N : P (CN — N, a, q, r, p) ? ? (—CN — + a) ? —CN —

Y ? (—c— + r) , ? (r)

(5)

c?CN r

q (1?p) where ? = 1?(1?p) r . We provide the complete derivation of equation 5, along with the conditional posterior distributions over r and p, in appendix A.2. Posterior inference for the NBNB model involves alternating between (a) sampling CN from P (CN — N, a, q, r, p) using the chaperones algorithm and (b) sampling r and p from their respective conditional posteriors using, e.g., slice sampling [17].

3.2

The NBD Model

Although ? = NegBin (a, q) will yield plausible values of K, ? = NegBin (r, p) may not be sufficiently flexible to capture realistic properties of N1 , . . . , NK , especially when K is large. For example, in a record-linkage application involving two otherwise noise-free databases containing thousands of records, K will be large and each Nk will be at most two. A negative binomial distribution cannot capture this property. We therefore define a second KP model?the negative binomial?Dirichlet (NBD) model?by taking a nonparametric approach to modeling N1 , . . . , NK and drawing ? from an infinite-dimensional Dirichlet distribution over the positive integers:

? = NegBin (a, q) and ? — ?, ?(0) ? Dir ?, ?(0) , (6) (0)

(0)

where ? ¿ 0 is a fixed concentration parameter and ?(0) = (?1 , ?2 , ? ? ?) is a fixed base measure P? (0) (0) with m=1 ?m = 1 and ?m ? 0 for all m. The probability of CN conditioned on N and ? is Y P (CN — N, a, q, ?) ? ? (—CN — + a) q —CN — —c—! ?—c— . (7) c?CN

Posterior inference for the NBD model involves alternating between (a) sampling CN from P (CN — N, a, q, ?) using the chaperones algorithm and (b) sampling ? from its conditional posterior:

(0) (0) ? — CN , ?, ?(0) ? Dir ? ?1 + L1 , ? ?2 + L2 , . . . , (8) where Lm is the number of clusters of size m in CN . Although ? is an infinite-dimensional vector, only the first N elements affect P (CN — a, q, ?). Therefore, it is sufficient to sample the PN (N + 1)-dimensional vector (?1 , . . . , ?N , 1 ? m=1 ?m) from equation 8, modified accordingly, and retain only ?1 , . . . , ?N . We provide complete derivations of equations 7 and 8 in appendix A.3. 3.3

The Microclustering Property for the NBNB and NBD Models

Figure 1 contains empirical evidence suggesting that the NBNB and NBD models both exhibit the microclustering property. For each model, we generated samples of MN / N for N = 100, . . . , 104 . For the NBNB model, we set a = 1, q = 0.5, r = 1, and p = 0.5 and generated the samples using rejection sampling. For the NBD model, we set a = 1, q = 0.5, and ? = 1 and set ?(0) to be a geometric distribution over N = {1, 2, . . .} with a parameter of 0.5. We generated the samples using MCMC methods. For both models, MN / N appears to converge to zero in probability as N ? ?, as desired. In appendix B, we also prove that a variant of the NBNB model exhibits the microclustering property. 4

3.4

Application to Entity Resolution

KP models can be used to perform entity resolution. In this context, the data points x1 , . . . , xN are observed records and the K clusters are latent entities. If each record consists of F categorical fields, then CN ? KP model ? f k — ?f , ? f ? Dir ?f , ? f

(9)

zn ? ?(CN , n) xf n — zn , ? f 1 , . . . , ? f K ? Cat (? f zn)

(10) (11) (12)

for f = 1, . . . , F , k = 1, . . . , K, and n = 1, . . . , N , where ?(CN , n) maps the nth record to a latent cluster assignment zn according to CN . We assume that ?f ¿ 0 is distributed as ?f ? Gam (1, 1), while ? f is fixed. Via Dirichlet?multinomial conjugacy, we can marginalize over ? 11 , . . . , ? F K to obtain a closed-form expression for P (x1 , . . . , xN — z1 , . . . , zN , ?f , ? f). Posterior inference involves alternating between (a) sampling CN from P (CN — x1 , . . . , xN , ?f) using the chaperones algorithm accompanied by appropriate likelihood terms, (b) sampling the parameters of the KP model from their conditional posteriors, and (c) sampling ?f from its conditional posterior using slice sampling.

4

Experiments

In this section, we compare two entity resolution models based on the NBNB model and the NBD model to two similar models based on the DP mixture model [10] and the PYP mixture model [11]. All four models use the likelihood in equations 10 and 12. For the NBNB model p and the NBD model, we set a and q to reflect a weakly informative prior belief that E[K] = Var[K] = N2 . For the NBNB model, we set ?r = sr = 1 and up = vp = 2.4 For the NBD model, we set ? = 1 and set ?(0) to be a geometric distribution over N = {1, 2, . . .} with a parameter of 0.5. This base measure reflects a prior belief that E[Nk] = 2. Finally, to ensure a fair comparison between the two different classes of model, we set the DP and PYP concentration parameters to reflect a prior belief that E[K] = N2 . We assess how well each model ?fits? four data sets typical of those arising in real-world entity resolution applications. For each data set, we consider four statistics: (a) the number of singleton clusters, (b) the maximum cluster size, (c) the mean cluster size, and (d) the 90th percentile of cluster sizes. We compare each statistic?s true value to its posterior distribution according to each of the models. For each model and data set combination, we also consider five entity-resolution summary statistics: (a) the posterior expected number of clusters, (b) the posterior standard error, (c) the false negative rate, (d) the false discovery rate, and (e) the posterior expected value of ?f = ? for f = 1, . . . , F . The false negative and false discovery rates are both invariant under permutations of 1, . . . , K [5, 18]. 4.1

Data Sets

We constructed four realistic data sets, each consisting of N records associated with K entities. Italy: We derived this data set from the Survey on Household Income and Wealth, conducted by the Bank of Italy every two years. There are nine categorical fields, including year of birth, employment status, and highest level of education attained. Ground truth is available via unique identifiers based upon social security numbers; roughly 74% of the clusters are singletons. We used the 2008 and 2010 databases from the Fruili region to create a record-linkage data set consisting of N = 789 records; each Nk is at most two. We discarded the records themselves, but preserved the number of fields, the empirical distribution of categories for each field, the number of clusters, and the cluster sizes. We then generated synthetic records using equations 10 and 12. We created three variants of this data set, corresponding to ? = 0.02, 0.05, 0.1. For all three, we used the empirical distribution of categories for field f as ? f . By generating synthetic records in this fashion, we preserve the pertinent characteristics of the original data, while making it easy to isolate the impacts of the different priors over partitions. NLTCS5000: We derived this data set from the National Long Term Care Survey (NLTCS)5 ?a longitudinal survey of older Americans, conducted roughly every six years. We used four of the 4 5

We used p ? Beta (2, 2) because a uniform prior implies an unrealistic prior belief that E[Nk] = ?. http://www.nltcs.aas.duke.edu/

5

available fields: date of birth, sex, state of residence, and regional office. We split date of birth into three separate fields: day, month, and year. Ground

truth is available via social security numbers; roughly 68% of the clusters are singletons. We used the 1982, 1989, and 1994 databases and down-sampled the records, preserving the proportion of clusters of each size and the maximum cluster size, to create a record-linkage data set of N = 5, 000 records; each Nk is at most three. We then generated synthetic records using the same approach that we used to create the Italy data set. Syria2000 and SyriaSizes: We constructed these data sets from data collected by four human-rights groups between 2011 and 2014 on people killed in the Syrian conflict [19, 20]. Hand-matched ground truth is available from the Human Rights Data Analysis Group. Because the records were hand matched, the data are noisy and potentially biased. Performing entity resolution is non-trivial because there are only three categorical fields: gender, governorate, and date of death. We split date of death, which is present for most records, into three separate fields: day, month, and year. However, because the records only span four years, the year field conveys little information. In addition, most records are male, and there are only fourteen governorates. We created the Syria2000 data set by down-sampling the records, preserving the proportion of clusters of each size, to create a data set of N = 2, 000 records; the maximum cluster size is five. We created the SyriaSizes data set by down-sampling the records, preserving some of the larger clusters (which necessarily contain withindatabase duplications), to create a data set of N = 6, 700 records; the maximum cluster size is ten. We provide the empirical distribution over cluster sizes for each data set in appendix D. We generated synthetic records for both data sets using the same approach that we used to create the Italy data set. 4.2

Results

We report the results of our experiments in table 1 and figure 2. The NBNB and NBD models outperformed the DP and PYP models for almost all variants of the Italy and NLTCS5000 data sets. In general, the NBD model performed the best of the four, and the differences between the models? performance grew as the value of ? increased. For the Syria2000 and SyriaSizes data sets, we see no consistent pattern to the models? abilities to recover the true values of the data-set statistics. Moreover, all four models had poor false negative rates, and false discovery rates?most likely because these data sets are extremely noisy and contain very few fields. We suspect that no entity resolution model would perform well for these data sets. For three of the four data sets, the exception being the Syria2000 data set, the DP model and the PYP model both greatly overestimated the number of clusters for larger values of ?. Taken together, these results suggest that the flexibility of the NBNB and NBD models make them more appropriate choices for most entity resolution applications.

5

Summary

Infinitely exchangeable clustering models assume that cluster sizes grow linearly with the size of the data set. Although this assumption is reasonable for some applications, it is inappropriate for others. For example, when entity resolution is treated as a clustering problem, the number of data points in each cluster should remain small, even for large data sets. Applications like this re-

quire models that yield clusters whose sizes grow sublinearly with the size of the data set. We introduced the microclustering property as one way to characterize models that address this requirement. We then introduced a highly flexible class of models?KP models?that can exhibit this property. We presented two models within this class?the NBNB model and the NBD model?and showed that they are better suited to entity resolution applications than two infinitely exchangeable clustering models. We therefore recommend KP models for applications where the size of each cluster should be unrelated to the size of the data set, and each cluster should contain a negligible fraction of the total number of data points. Acknowledgments We thank Tamara Broderick, David Dunson, Merlise Clyde, and Abel Rodriguez for conversations that helped form the ideas in this paper. In particular, Tamara Broderick played a key role in developing the idea of microclustering. We also thank the Human Rights Data Analysis Group for providing us with data. 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. 6

? ? ? ? ? ? ? ? ? ?
? ?
? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ?
? ?
?
?
?
?
?
? ?
?
?
?
?
?
? ? ?
? ? ?
? ? ? ?
?
? ? ?
? ? ?
? ?
?
?
?
?
?
? ?

?
?
1.35
?
? ?
? ? ?
? ?
?
? ?
? ?
? ?
? ?
? ?
?
?
?
?
?
? ? ? ? ?
7 1.30
? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ?
? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ?
? ?
? ?
? ? ? ? ? ? ? ? ? ?
? ? ? ? ?
? ? ? ?
? ?
? ?
? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ?
?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ? ? ? ? ?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
2.5

3.0
3.5
4.0
7
1.30
1.35
500
8
1.40
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
1.5
2.0
2.5
90th Percentile of Cluster Sizes
? ? ? ? ?
2.0
? ?
? ?
90th Percentile of Cluster Sizes
?
2.0
6
Mean Cluster Size 1.25
? ?
1.8
5
4
? ? ? ?
? ? ?
? ? ? ? ? ?
1.6
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
3
Maximum Cluster Size
? ? ? ? ? ? ? ? ? ? ?
1.4
1.58 1.60 1.62 1.64 1.66 1.68 1.70
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
2
450
?
1.2
9

? ? ? ? ? ? ? ? ?
1.0
8
? ?
? ? ? ?
90th Percentile of Cluster Sizes
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
7
Mean Cluster Size
?
? ? ? ? ? ? ? ?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ? ?
1.25
?
1.20
?
?
? ? ? ? ?
? ? ? ? ? ?
3.0
? ? ?
? ? ? ? ? ?
2.8
? ? ?
?
1.15
6
?
? ?
? ? ? ?
2.6
? ? ?
? ?
? ? ? ? ?
2.4
?
? ? ? ? ? ? ?
2.2
? ?
?
? ? ? ?
2.0
? ?

?
?
?
Mean Cluster Size
5
?
?
? ? ? ? ? ? ? ? ?
90th Percentile of Cluster Sizes
?
?
?
?
1.10
4
400
Singleton Clusters
?
?
?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ?
? ?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ? ? ? ? ?
?
1.8
? ?
Maximum Cluster Size
?
3
? ?
?
?
1.7
? ?
10
? ? ? ? ? ?

8
? ? ? ? ?
6
? ? ? ? ? ? ? ? ? ? ? ?
1.6
? ?
? ? ? ? ?
?
1.5
? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ?
?
Mean Cluster Size
? ?
?
4
? ? ? ? ? ?
?
Maximum Cluster Size
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
350
? ? ? ? ? ? ? ?
2
1800
?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ?
1.4
18
? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ?
16
1700
Singleton Clusters
? ?
14
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
1600
?
? ? ? ? ?
12

1600 ? ? ? ?
? ? ? ? ? ? ?
10
1400
? ? ? ?
8
1200
Singleton Clusters ? ? ?
? ? ? ? ? ? ?
6
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
1000
? ? ? ? ? ? ? ?
Maximum Cluster Size
3000
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
DP(0.02) PYP(0.02) NBNB(0.02) NDB(0.02) DP(0.05) PYP(0.05) NBNB(0.05) NDB(0.05) DP(0.1) PYP(0.1) NBNB(0.1) NDB(0.1)
2500
Singleton Clusters
2000
? ? ? ?
(a) Italy: NBD model ¿ NBNB model ¿ PYP mixture model ¿ DP mixture model.
(b) NLTCS5000: NBD model ¿ NBNB model ¿ PYP mixture model ¿ DP mixture model.
?
?
?
? ?
?
? ?
? ? ?
?
? ?
?
? ?
? ?
? ?
? ?
? ?
(c) Syria2000: the models perform similarly because there are so few fields.
? ? ?

? ? ?
? ? ?
? ? ?
? ? ?
? ? ?
? ? ?
?
?
?
?
?
?
?
?
?

(d) SyriaSizes: the models perform similarly because there are so few fields.

Figure 2: Box plots depicting the true value (dashed line) of each data-set statistic for each variant of each data set, as well as its posterior distribution according to each of the four entity resolution models.

Table 1: Entity-resolution summary statistics?the posterior expected number of clusters, the posterior standard error, the false negative rate (lower is better), the false discovery rate (lower is better), and the posterior expected value of ??for each variant of each data set and each of the four models. Data Set

True K

Variant

Model

E[K]

Std. Err.

FNR

FDR

E[?]

Italy

587

? = 0.02

DP PYP NBNB NBD

594.00 593.90 591.00 590.50

4.51 4.52 4.43 3.64

0.07 0.07 0.04 0.03

0.03 0.03 0.03 0.00

0.02 0.02 0.02 0.02

? = 0.05

DP PYP NBNB NBD

601.60 601.50 596.40 592.60

5.89 5.90 5.79 5.20

0.13 0.13 0.11 0.09

0.03 0.03 0.04 0.04
0.03 0.04 0.04 0.04
? = 0.1
DP PYP NBNB NBD
617.40 617.40 610.90 596.60
7.23 7.22 7.81 9.37
0.27 0.27 0.24 0.18
0.06 0.05 0.06 0.05
0.07 0.07 0.08 0.10
? = 0.02
DP PYP NBNB NBD
3021.70 3018.70 3037.80 3028.20
24.96 25.69 25.18 5.65
0.02 0.03 0.02 0.01
0.11 0.11 0.07 0.09
0.03 0.03 0.02 0.03
? = 0.05
DP PYP NBNB NBD
3024.00 3045.80 3040.90 3039.30
26.15 23.66 24.86 10.17
0.05 0.05 0.04 0.03
0.13 0.10 0.06 0.07
0.06 0.05 0.05 0.06
? = 0.1
DP PYP NBNB NBD
3130.50 3115.10 3067.30 3049.10
21.44 25.73 25.31 16.48
0.12 0.13 0.11 0.09
0.09 0.10 0.08 0.08
0.10 0.10 0.11 0.12
? = 0.02
DP PYP NBNB NBD
1695.20 1719.70 1726.80 1715.20
25.40 36.10 27.96 51.56
0.70 0.71 0.70 0.67
0.27 0.26 0.28 0.28
0.07 0.04 0.05 0.02
? = 0.05
DP PYP NBNB NBD
1701.80 1742.90 1738.30 1711.40
31.15 24.33 25.48 47.10
0.77 0.75 0.74 0.69
0.31 0.32 0.31 0.32
0.07 0.04 0.04 0.03
? = 0.1
DP PYP NBNB NBD

1678.10 1761.20 1779.40 1757.30
40.56 39.38 29.84 73.60
0.81 0.81 0.77 0.74
0.19 0.22 0.26 0.25
0.18 0.08 0.04 0.03
? = 0.02
DP PYP NBNB NBD
4175.70 4234.30 4108.70 3979.50
66.04 68.55 70.56 70.85
0.65 0.64 0.65 0.68
0.17 0.19 0.19 0.20
0.01 0.01 0.01 0.03
? = 0.05
DP PYP NBNB NBD
4260.00 4139.10 4047.10 3863.90
77.18 104.22 55.18 68.05
0.71 0.75 0.73 0.75
0.21 0.18 0.20 0.22
0.02 0.04 0.04 0.07
? = 0.1
DP PYP NBNB NBD
4507.40 4540.30 4400.60 4251.90
82.27 100.53 111.91 203.23
0.80 0.80 0.80 0.82
0.19 0.20 0.23 0.25
0.03 0.03 0.03 0.04
NLTCS5000
Syria2000
SyriaSizes
3,061
1,725
4,075
8

2 References

[1] P. Christen. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, 2012. [2] P. Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering, 24(9), 2012. [3] W. E. Winkler. Overview of record linkage and current research directions. Technical report, U.S. Bureau of the Census Statistical Research Division, 2006. [4] R. C. Steorts, R. Hall, and S. E. Fienberg. A Bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Society, In press. [5] R. C. Steorts. Entity resolution with empirically motivated priors. Bayesian

Analysis, 10(4):849? 875, 2015. [6] R. C. Steorts, R. Hall, and S. E. Fienberg. SMERED: A Bayesian approach to graphical record linkage and de-duplication. Journal of Machine Learning Research, 33:922?930, 2014. [7] T. Broderick and R. C. Steorts. Variational bayes for merging noisy databases. In NIPS 2014 Workshop on Advances in Variational Inference, 2014. arXiv:1410.4792. [8] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society Series B, pages 731?792, 1997. [9] J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. arXiv:1502.06241, 2015. [10] J. Sethuraman. A constructive definition of Dirichlet priors. Statistica Sinica, 4:639?650, 1994. [11] H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. Statistica Sinica, 13(4):1211?1236, 2003. [12] J. F. .C Kingman. The representation of partition structures. Journal of the London Mathematical Society, 2(2):374?380, 1978. [13] D. Aldous. Exchangeability and related topics. ?cole d??t? de Probabilit?s de Saint-Flour XIII?1983, pages 1?198, 1985. [14] H. M. Wallach, S. Jensen, L. Dicker, and K. A. Heller. An alternative prior process for nonparametric Bayesian clustering. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010. [15] V. F. Kolchin. A problem of the allocation of particles in cells and cycles of random permutations. Theory of Probability & Its Applications, 16(1):74?90, 1971. [16] J. Pitman. Combinatorial stochastic processes. ?cole d??t? de Probabilit?s de Saint-Flour XXXII?2002, 2006. [17] R. M. Neal. Slice sampling. Annals of Statistics, 31:705?767, 2003. [18] R. C. Steorts, S. L. Ventura, M. Sadinle, and S. E. Fienberg. A comparison of blocking methods for record linkage. In International Conference on Privacy in Statistical Databases, pages 253?268, 2014. [19] M. Price, J. Klingner, A. Qtiesh, and P. Ball. Updated statistical analysis of documentation of killings in the Syrian Arab Republic, 2013. United Nations Office of the UN High Commissioner for Human Rights. [20] M. Price, J. Klingner, A. Qtiesh, and P. Ball. Updated statistical analysis of documentation of killings in the Syrian Arab Republic. Human Rights Data Analysis Group, Geneva, 2014.

9