Learning Dirichlet Processes from Partially Observed Groups

Avinava Dubey*, Indrajit Bhattacharya†, Mrinal Das†, Tanveer Faruquie* and Chiranjib Bhattacharyya†
*IBM India Research Lab, {avinava.dubey, tanveer.faruquie}@gmail.com
†Indian Institute of Science, {indrajit,mrinal,chiru}@csa.iisc.ernet.in

***Abstract*—Motivated by the task of vernacular news analysis using known news topics from national news-papers, we study the task of topic analysis, where given source datasets with observed topics, data items from a target dataset need to be assigned to observed source topics or to new ones. Using Hierarchical Dirichlet Processes for addressing this task imposes unnecessary and often inappropriate generative assumptions on the observed source topics. In this paper, we explore Dirichlet Processes with partially observed groups (POG-DP). POG-DP avoids modeling the given source topics. Instead, it directly models the conditional distribution of the target data as a mixture of a Dirichlet Process and the posterior distribution of a Hierarchical Dirichlet Process with known groups and topics. This introduces coupling between selection probabilities of all topics within a source, leading to effective identification of source topics. We further improve on this with a Combinatorial Dirichlet Process with partially observed groups (POG-CDP) that captures finer grained coupling between related topics by choosing intersections between sources. We propose novel inference algorithms for these models using collapsed Gibbs sampling. We evaluate our models in three different real-world applications. Using extensive experimentation, we compare against several baselines to show that our model performs significantly better in all three applications.**

***Keywords*-topic analysis; grouped data; partial observations; Dirichlet Process**

I. Introduction

Many applications require analysis of a *target data collection* in the context of prior knowledge, specified through a *source data collection*. We use as our main motivation the task of vernacular news analysis. Suppose we need to identify news topics from a target collection of vernacular news stories. It may not often be possible to discover such news topics from scratch, given limited linguistic and other resources available for the vernacular language. Instead, we can start from parallel news stories from the vernacular news paper and one or more English or national language newspapers from the same geographical region or country, and use as prior knowledge the observed news topics from the national or English news papers. This task is meaningful, since being from the same geographical region, news topics are expected to be shared between the news papers, with each news-paper being likely to report on some additional news topics that would be of interest to the regional community. So we specify as our source collection the news stories from these different news-papers, along with their known news topics. Thus the vernacular news analysis task becomes one of *identifying* observed news topics from other newspapers, and additionally *discovering* new regional news stories.

This task topic analysis given sources with observed topics arises in other domains as well. Consider a company that performs customer service analysis. It receives customer feedback documents from different types of companies, and identifies the different issues or complaints mentioned by their customers. After having analyzed documents from some companies, a host of possible issues have already been identified. Now, given target data from a new company, instead of trying to discover issues from scratch, it is more worthwhile to identify known issues from a source collection consisting of known issues these previous companies, and discover new ones, if any.

Observe that this task is different from that of simultaneously discovering topics from multiple document collections. The Hierarchical Dirichlet Process (HDP) [1] was proposed as an extension of Dirichlet Processes [2], [3] for non-parametric clustering of multiple groups of data. The HDP describes a mixture model for every group, and additionally allows mixture components to be shared across groups. In our task, we are *given* the mixture components, or topics, in some of the data groups, which we call sources, and we are interested in discovering whether data in a new target dataset shares components, or topics, from existing groups.

This is often a significant difference in practice. The generative process that the HDP, or any multi-task generative model, assumes for all the groups may not always be appropriate. For example, the observed topics in the existing groups, or sources, may have been identified by human experts using complex background knowledge. More importantly, it is not necessary to model the generative process of the known source components, when they are given, and we are only interested in the components for the target dataset.

In this paper, we propose the Dirichlet Process with Partially Observed Groups (POG-DP) for the task of topic analysis in a target dataset, using the knowledge of observed topics in one or more source datasets. Like the mixture of Dirichlet Processes [3] and the Dependent Dirichlet Processes [4], [5], [6], the POG-DP introduces dependences between multiple Dirichlet Processes. However, the POG-DP *does not model* the source topics. Instead, it directly models the conditional distribution of the target data directly,

as a mixture of the posterior distribution of an HDP with known groups and topics, and a Dirichlet Process. For generating each target data instance, the process randomly chooses either an existing group or source followed by an existing topic in that source, or a new group different from the sources followed by an existing or new topic in that group. Indeed, our experimental results show that POG-DP outperforms models such as the HDP that makes inappropriate generative assumptions on the observed source topics.

We further propose the Combinatorial Dirichlet Process that selects topics from *group intersections*. In the Dirichlet Process and the Hierarchical Dirichlet Process with observed groups, the posterior selection probabilities of different topics are decoupled. The POG-DP improves identification of source topics by coupling together the selection probabilities of all topics within a group. However, this is inappropriate when *some* topics within a group are related, but the range of topics in a group is still diverse. When different groups have overlaps among their topics, the overlaps represent coherent subsets of topics within groups. For example, the topics at the intersection of a sports news paper and regional news paper would represent sports popular in that region. Given sources with overlapping topics, the combinatorial Dirichlet Process with partially observed groups (POG-CDP) introduces coupling only among topics in group intersections, thereby improving over the POG-DP.

We propose efficient inference algorithms based on collapsed Gibbs sampling for the proposed models. We evaluate the models in three different applications, and show that they significantly outperform various baselines in identifying known topics and discovering new ones.

The main contributions of this paper are the following. (a) Motivated by the real-life problem of analyzing news corpora using prior knowledge of source topics, we explore the Dirichlet Process with partially observed groups (POG-DP), that model the conditional distribution of a target dataset, given multiple sources with known topics, improving over the HDP model. (b) For sources with overlapping topics, we propose the Combinatorial Dirichlet Process with partially observed groups (POG-CDP), that efficiently represents and learn the target relevance of arbitrary subsets of overlapping sources. (c) We propose Gibbs sampling based efficient inference strategies for our models. (d) We use the proposed models to address three different real life tasks — vernacular news analysis, customer satisfaction analysis and news group analysis — and show that they significantly outperform various baselines.

The rest of this paper is organized as follows. In Section III, we begin with a review of the Dirichlet Process and the Hierarchical Dirichlet Process, and then motivate the Dirichlet Process with partially observed groups, and its combinatorial counterpart. Inference algorithms for the model, based on collapsed Gibbs sampling, are presented in Section IV, experimental evaluation in Section V, and we conclude in Section VI.

II. Related Work

The Dirichlet Process [2] is a popular non-parametric Bayesian prior over distributions, and the Dirichlet Process mixture model [2], [3] is used extensively in clustering applications having a single collection of data items, where the number of mixture components is unknown. The Hierarchical Dirichlet Process (HDP) extends the Dirichlet Process (DP) for simultaneous clustering of multiple collections or groups of data items, where in addition to finding mixture components for clustering data in each group, we require the components to be shared across the groups. The HDP uses a DP G_j for each group, where all of them are drawn independently and identically from a base measure G_0. For mixture components to be shared across groups with non-zero probability, G_0 needs to be discrete. The HDP models G_0 as a Dirichlet Process as well, resulting in a two-level hierarchy of Dirichlet Processes.

The HDP introduces dependence between multiple Dirichlet Processes. Other models have been proposed to introduce dependence between Dirichlet Processes, such as the mixture of Dirichlet Processes [3], and the various dependent Dirichlet Processes [4], [5], [6]. However, the HDP is most suited for the purpose of sharing mixture components across discrete groups. It can be shown to be a special case of Dependent Dirichlet Process model of MacEachern et.al. [5], [6], where instead of single mixture components, there are indexed collections of related mixture components.

The Hierarchical Dirichlet Process and the Dependent Dirichlet Process introduce dependence for clustering grouped data, *when the groups are observed*. The Nested Dirichlet Process [7] addresses a related problem of clustering the groups according to similarities between them, but still assumes the group structure of the data to be known.

The HDP-HMM model [1], [8] extends the Hierarchical Dirichlet Process to handle completely unobserved group variables for the task of learning Hidden Markov Models with unknown number of hidden states. In this paper, we explore models for dependent Dirichlet Processes for *partially observed* groups and topics. The POG-DP model is a mixture of a Dirichlet Process and the posterior distribution of an HDP with known groups and topics. To the best of our knowledge, our work is the first exploration of coupled DP models for data with partially observed groups and topics.

Differently from HDP-HMMs that parameterize choices over groups, the combinatorial Dirichlet Process parameterizes group intersections. This introduces finer grained coupling between components within groups. While it is possible and meaningful to investigate combinatorial HDPs for data with completely observed and completely unobserved groups as well, for the specific task of topic analysis

with observed source topics, we only combinatorial DPs with partially observed groups in this paper.

[mrinal]Recently [9] proposed an approach to implement topic models over sub-corpus instead of the full large dataset, and then to combine the topics. This enables to apply topic models in a distributed and incremental framework. The major difference with our model is that the ensemble approach treats each sub-corpus as independent of each other, and even in the incremental setting existing topics do not influence topics to be detected from the new sub-corpus. Whereis in our case the approach learns Dirichlet Processes conditioned on the existing observed topics and groups.[/mrinal]

III. Generative Topics Models for Grouped Data

In this section, we first briefly review the Dirichlet Process (DP)[2], [3] for modeling topics in a single group, and the hierarchical Dirichlet Process (HDP) [1] for multiple observed groups, before discussing models for partially observed group variables.

A. Completely Observed Groups

Consider a measurable space $(\Theta, \mathcal{B})$, where $\Theta \subset \mathbb{R}^d$, $\mathcal{B}$ is the Borel σ-algebra of subsets of $\mathbb{R}^d$. Let G_0 be a probability measure over $(\Theta, \mathcal{B})$, and α be a positive real number. A Dirichlet Process $DP(\alpha, G_0)$ is a distribution over probability measures, with α as a strength parameter, and G_0 as a base distribution. A random probability measure G over $(\Theta, \mathcal{B})$ is distributed according to $DP(\alpha, G_0)$, if for any finite partition $(A_1, \ldots, A_k)$ of Θ, $(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$, where $Dir(\alpha_1, \ldots, \alpha_k)$ is a finite dimensional Dirichlet distribution.

The Dirichlet Process (DP) can be used for defining non-parametric mixture models for a single data collection as follows. Consider a collection $\mathcal{D}$ of data items of the form $\{x_i, \eta_i\}$, where η_i denotes the mixture component from which data item x_i is generated. The generative process first samples each η_i, independently and identically (i.i.d.) from a Dirichlet Process, instead of a finite dimensional distribution, and then samples x_i i.i.d. using η_i.

1. $G \sim DP(\alpha, G_0)$
2. For the i^{th} data instance
3. $\quad \eta_i \sim G$
4. $\quad x_i \sim F(\eta_i)$

For the task of discovering topics from documents, for example for discovering news topics from a collection of news articles, each mixture component η_i corresponds to the topic from which the words of the document x_i are generated. So, for this paper, *we use components and topics interchangeably*. Also, we consider the generating distribution $F()$ for each data item to be a multinomial, and the base distribution G_0 of the DP to be a Dirichlet $Dir(\lambda)$ for conjugacy. Further, since x_i is vector valued, each element of x_i is generated i.i.d from $Mult(\eta_i)$. We will use this process for $x_i \sim F(\eta_i)$ for all of the models that we present in this paper, without describing it explicitly. (Note that here each document corresponds to a single topic, unlike the Latent Dirichlet Allocation model [10])

Following this generation process, the joint distribution over the data and the mixture components looks as follows: $P(\{x_i, \eta_i\}; \alpha, G_0) = P(\{\eta_i\}; \alpha, G_0) \prod_i P(x_i|\eta_i)$. On integrating G out, the conditional distribution for the n^{th} draw $\eta_n \sim G$ from a Dirichlet Process, given the previous $n-1$ draws $\eta_{1:n-1}$ is given by

$$\begin{aligned} & \eta_n | \eta_1 \ldots \eta_{n-1}; \alpha, G_0 \\ \sim \quad & \tfrac{\alpha}{n-1+\alpha} G_0 + \textstyle\sum_{i=1}^{K} \tfrac{m_i}{n-1+\alpha} \delta_{\phi_i} \end{aligned} \tag{1}$$

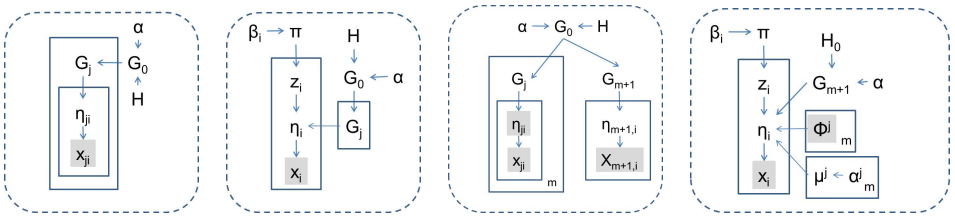
where $\phi_1 \ldots \phi_K$ are the unique values taken by $\eta_1 \ldots \eta_{n-1}$ with corresponding counts $m_1 \ldots m_K$ [11]. Using this joint distribution, the task is usually to infer the unknown mixture components η_i from the observed data x_i.

The Dirichlet Process mixture model is useful for discovering mixture components from a single collection of data items. However, in many applications, such as in the case of analyzing multiple news corpora, data may be partitioned into groups. The task then would be to discover mixture components within each group. Additionally, the task may require the mixture components to be shared among the different groups. For our news example, we typically do not want all the news topics in the different news corpora to be distinct, if they are all reporting news from the same geographic region or country. The Hierarchical Dirichlet Process (HDP) [1] extends the DP for grouped data of the form $\mathcal{D} = \{x_i, \eta_i, z_i\}$, where η_i represents the mixture component for x_i as before, and z_i represents the group to which data item x_i belongs. If the partition of the data items into groups is completely known, then all the z_i variables are observed, so that $\mathcal{D}$ can be equivalently represented in grouped form as $\{x_{ji}, \eta_{ji}\}$, where x_{ji} denotes the i^{th} data item for the j^{th} group, and η_{ji} denotes the mixture component that generated x_{ji}. The HDP models the data from each group as coming from a non-parametric mixture model, so each η_{ji} is chosen i.i.d. from a Dirichlet Process G_j for each group, and in turn G_j's are chosen i.i.d. from a base probability measure G_0. To enable sharing of components across groups, G_0 needs to be a discrete distribution. Specifically, G_0 is modeled as a Dirichlet process $DP(\alpha, H)$. The complete HDP generative process is given as follows:

1. $G_0 \sim DP(\gamma, H)$
2. For each group j
3. $\quad G_j \sim DP(\alpha, G_0)$
4. $\quad$ For each item i in group j:
5. $\quad\quad \eta_{ji} \sim G_j$
6. $\quad\quad x_{ji} \sim F(\eta_{ji})$

This is shown using the plate notation in Figure 1(a).

The problem addressed by the HDP is inferring the unobserved mixture components η_{ji} for each data item, using the observed data items x_{ji} partitioned into groups. On



(a) (b) (c) (d)

Figure 1. Plate representation for (a) HDP, (b) HDP with completely unobserved groups, (c) PO-HDP and (b) POG-DP. Observed variables are shaded.

integrating out the G_j's and G_0, the conditional distribution for the n^{th} draw from the j^{th} group η_{jn} given all previous draws from group j and other groups turns out to be:

$$\eta_{jn} \mid \eta_{1:j-1}, \eta_{j1} \ldots \eta_{j,n-1}; \alpha, H \sim \sum_i \frac{n_i^j}{n-1+\alpha}\delta_{\theta_i^j} + \frac{\alpha}{n-1+\alpha}\left[\sum_k \frac{m_k}{m_\cdot+\gamma}\delta_{\psi_k} + \frac{\gamma}{m_\cdot+\gamma}H\right] \quad (2)$$

where $\{\theta_i^j\}$ are the draws made by the j^{th} group from G_0 with corresponding counts $\{n_i^j\}$, and $\{\psi_k\}$ are the unique draws made by all the groups from H with corresponding counts $\{m_k\}$, $m_k = \sum_{i,j}\delta(\theta_i^j, \psi_k)$ and $m_\cdot = \sum_k m_k$ [1]. Note that the topics $\{\theta_i^j\}$ generated within the i^{th} group are *not unique* for the HDP.

B. Completely Unobserved Group and Topics

We first look at the most general learning problem for grouped data, where the groups z_i, in addition to the mixture components η_i are unobserved for *all* x_i, before considering partially known groups and topics. Imagine that we are given a large collection of news stories from different news papers. The task is to identify the news topics, as well as identifying the news papers, corresponding to each news article.

Though the case of completely unobserved groups has been studied in the HDP-HMM model where a sequential structure is additionally observed over the words [1], [8], we are not aware of work on learning general HDP with completely unobserved groups. Here we describe the generative process and conditional distributions for it, which we then extend for the partially observed case in Subsection III-C.

In this scenario, $\{x_{ji}\}$ cannot be taken as the representation of the observed data, and the group-wise generative process is not possible. Instead, we have a sequential generative process over the data items, where, for the i^{th} data item, z_i is first sampled from a random distribution π, followed by generation of η_i, and then x_i. In the most general setting, the number of groups, or the number of news papers corresponding to the news stories in our example, may also be unknown. So we model z_i as a discrete random variable taking value from the set of positive integers $\mathbb{I}^+$. The prior distribution π then needs to be a probability distribution over $\mathbb{I}^+$. As such, we can model π as sampled from $GEM(\beta)$, defined using the stick breaking construction [12] as $\pi_k = \pi'_k \prod_{i=1}^{k-1}\pi'_i$, with $\pi'_i \sim Beta(1, \beta)$.

The complete generative process for HDP with unobserved groups looks as follows:

1. $G_0 \sim DP(\gamma, H)$
2. For each group j $j = 1 \ldots \infty$
3. $\quad G_j \sim DP(\alpha, G_0)$
4. $\pi \sim GEM(\beta)$
5. For each item i
6. $\quad z_i \sim Mult(\pi)$
7. $\quad \eta_i \sim G_{z_i}$
8. $\quad x_i \sim F(\eta_i)$

This is shown using the plate notation in Figure 1(b). This is similar to the generation process for the HDP-HMM model [1], without an additional sequential dependence over the group variables.

The task is to infer the latent mixture components η_i as well as the latent group variables z_i for each data item x_i. On integrating out G_0, G_j's and π, the conditional distribution for the component η_n corresponding to the n^{th} data item, given the components and groups of the previous data items, looks as follows:

$$\begin{aligned}\eta_n \mid \eta_1 \ldots \eta_{n-1}, z_1 \ldots z_n; \gamma, H, \alpha, \beta \sim & \sum_{k=1}^m \frac{n^k}{n-1+\beta}[\sum_i \frac{n_i^k}{n_\cdot^k+\alpha}\delta_{\theta_i^k} \\ & + \frac{\alpha}{n_\cdot^k+\alpha}\left(\sum_t \frac{m_t}{m_\cdot+\gamma}\delta_{\psi_t} + \frac{\gamma}{m_\cdot+\gamma}H\right)] \\ & + \frac{\beta}{n-1+\beta}\left[\sum_t \frac{m_t}{m_\cdot+\gamma}\delta_{\psi_t} + \frac{\gamma}{m_\cdot+\gamma}H\right] \end{aligned} \quad (3)$$

where the first term captures the probability of selecting one of the m exiting groups, and the second term that of selecting a new group. $\{\theta_i^k\}$, $\{n_i^k\}$, $\{\psi_t\}$ and $\{m_t\}$ are defined as for the HDP.

In the special case when the number of groups m is known, we can model z_i as coming from an m-dimensional Multinomial distribution with parameters π: $z_i \sim Mult(\pi)$. This is in turn sampled from a prior Dirichlet distribution $Dir(\beta)$. The corresponding conditional distribution looks as

follows:

$$\begin{aligned} & \eta_n \mid \eta_1 \dots \eta_{n-1}; \gamma, H, \alpha, \beta \\ \sim \quad & \sum_{k=1}^{m} \frac{\beta + n^k}{n-1+M\beta} [\sum_i \frac{n_i^k}{n^k + \alpha} \delta_{\theta_i^k} \\ & + \frac{\alpha}{n^k + \alpha} \left(\sum_t \frac{m_t}{m_\cdot + \gamma} \delta_{\psi_t} + \frac{\gamma}{m_\cdot + \gamma} H \right)] \end{aligned} \tag{4}$$

This specific model is not relevant for our task, and we do not include it in our evaluations for this paper. We next extend it for partially observed groups.

C. Partially Observed Groups

In the task that we focus on, we are given news stories from multiple source news-papers with identified news topics. For an additional target collection of news stories, we need to infer if they correspond to known news topics from existing news-papers, or to new topics. Assuming additional news topics cannot be added to the given news-papers, we can reformulate the question as follows: *"Which news stories in the target dataset correspond to existing news topics from the known news-papers, and which ones correspond to new news topics from some other unseen news-paper?"*

Formally, we represent the data $\mathcal{D}$ as being partitioned into two non-overlapping sets $\mathcal{D}_o = \{x_i^o, \eta_i^o, z_i^o\}$ where all the group variables $\{z_i^o\}$ and component (topic) variables $\{\eta_i^o\}$ are observed along with $\{x_i^o\}$, and $\mathcal{D}_u = \{x_i^u, \eta_i^u, z_i^u\}$, where $\{x_i^u\}$ are the observed variables, but $\{\eta_i^u\}$ and $\{z_i^u\}$ are unobserved. We refer to $\mathcal{D}_o$ as the **source data**, and to $\mathcal{D}_u$ as the **target data**. Also, let $1 \dots m$ denote the unique values taken by the observed group variables z_i^o in $\mathcal{D}_o$. We refer to these **existing groups as sources**. Given this source data $\mathcal{D}_o$ with observed groups and topics, the task then is to infer the unobserved group variables z_i^u and component variables η_i^u in the target data $\mathcal{D}_u$. Note that each unobserved topic variable η_i^u can take values from any of the existing source topics $\{\psi_k\}$, or some new topic. Similarly, z_i^u can take any of the known values $1 \dots m$ from $\mathcal{D}_o$, indicating that x_i^u corresponds to one of the sources, or some new values, indicating that x_i^u does not correspond to any of the sources.

In general, target documents in $\mathcal{D}_u$ may correspond to more than one new group. Further, for any new topic that is observed, all existing and new groups may be allowed to share this new topic. For our specific task, we restrict the scope of the problem. First, since we are interested only in determining if a new document in $\mathcal{D}_u$ corresponds to any of the sources or not, we only allow for one additional group in $\mathcal{D}_u$ corresponding to $z_i^u = m+1$. Further, we assume that no new topics may be added to given sources. So, any new topic that is created, has non-zero probability only under the new group $m+1$, and is not shared by any of the existing groups $1 \dots m$.

Before describing the generative process for $\mathcal{D}_u$, one last issue is left to be addressed. The new group $m+1$ must have its own distribution G_{m+1} over mixture components $\{\eta_i^u\}$. We define G_{m+1} as a Dirichlet Process, as for the other groups, to allow as many new topics as required. We have two choices for the base distribution for this DP. It could be the same base distribution G_0 used for $G_1 \dots G_m$, or it could be different distribution H_0. This leads to two different models for partially observed groups.

1) Partially observed HDP: When G_{m+1} is drawn from the same distribution as G_j $j = 1 \dots m$, then the generative process corresponds to a HDP where the groups $1 \dots m$ and the topics η^{ji} for data from these m groups are observed in $\mathcal{D}_o$, while the topics $\eta^{m+1,i}$ are unobserved for data from the new group $m+1$ in $\mathcal{D}_u$. We call this the Partially observed Hierarchical Dirichlet Process (PO-HDP). The generative process for $\mathcal{D}_u$ then assumes that data from $\mathcal{D}_o$ has already been generated by the HDP resulting in the observed topics $\eta_i^k, k = 1 \dots m$, and then proceeds with the HDP generation process for $\mathcal{D}_u$. All of the data points in $\mathcal{D}_u$ are assumed to come from group G_{m+1}. This is shown using the plate notation in Figure 1(c).

The conditional distribution for the n^{th} target data item in $\mathcal{D}_u$ given the previous $n-1$ draws for the PO-HDP is given by Equation (2), where m_k and n_i^j now denote the total counts over all data items in $\mathcal{D}_o$ and the first $n-1$ items in $\mathcal{D}_u$. However, note that the draws from G_0 $\{\theta_i^j\}$ for groups $1 \dots m$ are not observed in $\mathcal{D}_o$, and need to be inferred. We consider the PO-HDP model as a baseline for learning from partially observed groups.

2) Dirichlet Process with Partially Observed Groups: The alternative prior for G_{m+1} is to choose $H_0 \neq G_0$. This leads to our proposed model, which we call the Dirichlet Process with Partially Observed Groups (POG-DP).

We make the following observations for POG-DP:
(a) For $H_0 \neq G_0$, the target data $\mathcal{D}_u$ becomes conditionally independent of G_j $j = 1 \dots m$, as well as G_0, given the samples $\{\eta_i^o\}$ from $G_1 \dots G_m$. This means that we can model $\mathcal{D}_u$ without making any generative assumptions on $\{\theta_i^k\}, k = 1 \dots m$. This is an attractive modeling option, because in many scenarios, the labels in $\mathcal{D}_o$ may be generated by processes for which the HDP, or any other prior distribution, may not be appropriate. For example, the news topics for the known news papers may be curated by human experts using complex background knowledge and various linguistic and other resources. [mrinal] Topics are defined to be distribution over the words in the given vocabulary. However, topics derived by experts may not be in that form, rather will be in simple form as set of words which can be converted to a distribution by giving high value for specified words and low value for other words. Alternatively given a set of documents corresponding to a topic, the topic can be derived by simple word frequency information.[/mrinal]
(b) A consequence of sampling G_{m+1} from $H_0 \neq G_0$ is that all draws from G_{m+1} in $\mathcal{D}_o$ will be distinct from the known source topics $\{\psi_k\}$. In other words, no existing topics can be shared by the new group. This is not restrictive for

our application, where we only need to infer if a new news article corresponds to a new topic or an existing source topic. Thus there is no additional merit in the new group sharing topics with the sources.

Finally, we come to the generative process in the POG-DP model for target data $\mathcal{D}_u$ given source data $\mathcal{D}_o$. For the k^{th} existing group, or source, $(k = 1 \dots m)$, let $\{\phi_i^k\}$ be the *unique values* taken by $\{\eta_i : z_i = k\}$ with corresponding counts $\{n_i^k\}$. These represent the known, or given, topics in each source. Note that these are different from the draws $\{\theta_i^k\}$ for the k^{th} group in the PO-HDP, which need not be unique. We define $\tilde{G}_{m+1} \equiv G_{m+1}$, and $\tilde{G}_k \equiv \sum_i \mu_i^k \delta_{\phi_i^k}$, $(1 \leq k \leq m)$ is a multinomial distribution with parameter μ^k over the known topics ϕ^k for the k^{th} source. The multinomial parameter μ^k is in turn drawn from a Dirichlet distribution $Dir(\alpha^k)$ for each source.

1. $\mu^k \sim Dir(\alpha^k)$, $k = 1 \dots m$
2. $G_{m+1} \sim DP(\alpha^{m+1}, H_0)$
3. $\pi \sim Dir(\beta)$
4. For the i^{th} data item in $\mathcal{D}_u$
5. $\quad z_i^u \sim Mult(\pi)$
6. $\quad \eta_i^u \sim \tilde{G}_{z_i^u}$
7. $\quad x_i^u \sim F(\eta_i^u)$

This is shown using plate notation in Figure 1(d). Note that we *do not* define $\tilde{G}_k$ as the empirical distribution $\sum_i \frac{n_i^k}{n_\cdot^k} \delta_{\phi_i^k}$ using the counts from $\mathcal{D}_o$ as in the PO-HDP (Equation (2)). In the POG-DP, the target documents are free to have a different preference over topics inside a source, and this is learnt from the target data.

Also, note that H_0 does not need to be a discrete distribution any more like G_0. As such, we model H_0 as a finite dimensional Dirichlet distribution $Dir(\lambda)$. The conditional distribution $\eta_n^u | \eta^o, \eta_{1:n-1}^u, z_{1:n-1}^u, \alpha, \beta, H_0$ for the n^{th} target data item in $\mathcal{D}_u$ given the previous $n-1$ draws is now:

$$\begin{aligned} & \eta_n^u | \eta^o, \eta_{1:n-1}^u, z_{1:n-1}^u, \alpha, \beta, H_0 \\ \sim \; & \frac{\beta + n_\cdot^{m+1}}{m\beta + n - 1} \Big(\sum_{i=1}^{K^{m+1}} \frac{n_i^{m+1}}{\alpha^{m+1} + n_\cdot^{m+1}} \delta_{\phi_i^{m+1}} + \frac{\alpha^{m+1}}{\alpha^{m+1} + n_\cdot^{m+1}} H_0 \Big) \\ & + \sum_{k=1}^{m} \frac{\beta + n^k}{m\beta + n - 1} \left(\sum_{i=1}^{K^k} \frac{\alpha^i + n_i^k}{K^k \alpha^i + n^k} \delta_{\phi_i^k} \right) \end{aligned} \tag{5}$$

where $\{\phi_i^{m+1}\}$ are the *unique* draws in $\mathcal{D}_u$ from the new group distribution G_{m+1} with corresponding counts $\{n_i^{m+1}\}$, and $\{n_i^j\}$, $j = 1 \dots m$ are the number of draws in $\mathcal{D}_u$ from *unique* known topics $\{\phi_i^j\}$, $j = 1 \dots m$ from the m sources.

It is interesting to observe the differences with the PO-HDP conditional in Equation (2). For the PO-HDP, the selection of a specific topic θ_i^k boosts the posterior selection probability of only that topic through the selection count n_i^k (first term on RHS). However, for the POG-DP, the selection of a topic ϕ_i^k increases not only count n_i^k, but also the count $n_\cdot^k$ for the group. This increases the posterior selection probability of *all topics* in that group. Thus the posterior selection probabilities of all topics within a group are coupled in the POG-DP, unlike PO-HDP.

D. Modeling Group Intersections

As explained above, the posterior distribution over topics in the POG-DP couples together selection probabilities of all topics within the same group. Intuitively, this means than when *some news topics* from a specific news-paper are chosen many times for articles in the target data, the posterior selection probability of *all news topics* in that news-paper increases for the target data. When all news topics in a news-paper are related, for example in the case of a business news paper, such coupling is appropriate. Indeed, we will see in our experimental results that this helps to improve performance over the PO-HDP and variants of the Dirichlet Process where no such coupling occurs. However, when a news paper contains articles on diverse topics, coupling all topics within a news-paper is unreasonable.

Ideally, we would like to couple subsets of topics within a group that are statistically related. Unfortunately, searching over arbitrary subsets of topics within a group is not computationally feasible. However, given source data with more than one observed group or source, often different sources share topics, and source intersections have natural interpretations, and can be assumed to be related. For example, the topics at the intersection of a general news-paper for a region, and a sports news-paper, would relate to sports news for that region. Then we can introduce dependencies between topics that appear in *group intersections*, rather than in individual groups, or sources. In our second proposed model, which we call the combinatorial Dirichlet Process (CDP), we represent intersections by indexing arbitrary combinations of sources, and introduce selection probabilities for such intersections. Compared with the PO-HDP, where the a data item is chosen by first selecting a source, the generative process for the CDP first chooses a combination of sources, and then selects a topic from the intersection of the selected sources.

While it is meaningful to explore combinatorial DP with completely observed as well as completely unobserved group variables, in this paper we focus on our specific task use CDPs with partially observed groups ($\mathcal{D}_o$ and $\mathcal{D}_u$). Formally, for the partially observed combinatorial DP (POG-CDP), our representation of $\mathcal{D}_o$ remains the same as before, with z_i^o taking m unique values, indicating the sources. In $\mathcal{D}_u$, we now represent z_i^u as a binary vector valued random variable, with dimension m. z_i^u is now represents a subset of the m sources whose corresponding entries in z_i^u are 1. Define $\phi^{z_i^u} = \cap_{\{j : z_{ij}^u = 1\}} \phi^j$ as the set of shared components in known groups or sources for which z_i^u has 1. Then, given z_i^u, η_i would be sampled from $\tilde{G}_{z_i^u} \equiv \sum_t \mu_t^{z_i^u} \delta_{\phi_t^{z_i^u}}$. When z_i^u is $\mathbf{0}$, η_i^u is chosen from G_{m+1} corresponding to the new group $m+1$. Empty subsets can be handled by creating a dummy component ϕ_d and associating it with all empty subsets with $\mu_d = 1.0$.

To define the prior distribution over z_i^u, we assume that

groups are chosen independently for the intersection, and parameterize each element of z_i^u independently: $z_{ij}^u \sim Ber(\rho_j)$. The generative process for $\mathcal{D}_u$ in POG-CDP model looks similar to POG-DP, with the only difference being in the generation of z_i^u and in definition of $\tilde{G}_{z_i^u}$.

1. $\mu^k \sim Dir(\alpha)$, $k = 1 \dots M$
2. $G_{m+1} \sim DP(\alpha^{m+1}, H_0)$
3. For the i^{th} data item in $\mathcal{D}_u$
4. $\quad z_{ij}^u \sim Ber(\rho_j)$, $j = 1 \dots m$
5. $\quad \eta_i^u \sim \tilde{G}_{z_i^u}$
6. $\quad x_i^u \sim F(\eta_i^u)$

The conditional distribution for η_n^u given $\eta_{1:n-1}^u$ for the POG-CDP is also very similar — instead of one term per source as in the POG-DP, it has one term per source intersection z:

$$\begin{aligned} & \eta_n^u | \eta^o, \eta_{1:n-1}^u, z_{1:n-1}^u, \rho, H_0 \\ \sim \; & \frac{\rho(0)+n_\cdot^{m+1}}{\sum_i \rho(i)+n-1} \Big(\sum_{i=1}^{K^{m+1}} \frac{n_i^{m+1}}{\alpha^{m+1}+n_\cdot^{m+1}} \delta_{\phi_i^{m+1}} \\ & + \frac{\alpha^{m+1}}{\alpha^{m+1}+n_\cdot^{m+1}} H_0 \Big) \\ & + \sum_z \frac{\rho(z)+n_\cdot^z}{\sum_i \rho(i)+n-1} \left(\sum_{i=1}^{K^z} \frac{\alpha^z+n_i^z}{K^z \alpha^z + n_\cdot^z} \delta_{\phi_i^z} \right) \end{aligned} \quad (6)$$

Importantly, observe that only selection probabilities for components within source intersections are now coupled, instead of all components within the same source.

Additionally, the POG-CDP also enables us to answer richer queries compared to the POG-DP. The PO-DP model allows us to determine, given a target collection of news stories, if its articles correspond to existing topics from any of the known news-papers, or if it is from a new topic from a new news-paper. Using the PO-CDP, we can also ask if it comes from a topic *shared by some known news-papers*.

IV. Inference

Inference for the DP with partially observed groups (POG-DP) and combinatorial DP with partially observed groups (POG-CDP) involves determining the η_i^u and z_i^u values for each target data instance x_i^u, that maximizes the conditional likelihood $P(x^u | \eta; \beta, \alpha, \lambda)$ given the components $\phi = \{\phi^1, \dots, \phi^m\}$ of the m observed groups and the hyper-parameters α^m, β and λ.

In general, exact inference in infeasible, and we propose collapsed Gibbs sampling based inference algorithms for the proposed models. The algorithms for the two models follow the general sampling scheme, where new values for each of the random variables, z_i and η_i are sampled from their conditional distributions, given the current values of all other random variables:

$$\begin{aligned} & p(\eta_i | x, \eta_{-i}, z, \phi; \beta, \lambda, \alpha^m) \\ \propto \; & p(\eta_{i^*} | \eta_{-i^*}, z_{i^*}, z_{-i^*}; \alpha^m) p(X_{i^*} | \eta_{i^*}, x_{-i^*}, \phi, \lambda) \quad (7) \\ & p(z_i | x, \eta, z_{-i}, \phi; \beta, \lambda, \alpha^m) \\ \propto \; & p(z_{i^*} | z_{-i^*}; \beta) p(\eta_{i^*} | \eta_{-i^*}, z_{i^*}, z_{-i^*}; \alpha) \quad (8) \end{aligned}$$

The inference algorithm repeatedly samples η_i and z_i cyclically over data instances x_i using these conditional distributions until convergence.

When $z_i = m+1$, the data items are assigned using the Dirichlet Process mixture model $DP(\alpha^{m+1}, H_0)$. Using the partition structure of Dirichlet Process [13], [14], the posterior for η_i turns out to be

$$\begin{aligned} & p(\eta_i = k | \eta_{-i}, z_i = m+1, z_{-i}; \alpha^{m+1}) \\ = \; & \frac{n_{k,-i}^{m+1}}{\alpha^{m+1} + \sum_{k'} n_{k',-i}^{m+1}} \quad \text{if } k \text{ seen before} \\ = \; & \frac{\alpha^{m+1}}{\alpha^{m+1} + \sum_{k'} n_{k',-i}^{m+1}} \quad \text{if } k \text{ not seen before} \end{aligned}$$

where $n_{k,-i}^j$ denotes the count of the number of data items $x_{i'}^u$, excluding the i^{th} one, that have $\eta_{i'}^u = k$ and $z_{i'}^u = j$.

However, when z_i takes values from $1 \dots m$, data items are selected from one of the existing groups. Then, integration out π, we get

$$\begin{aligned} & p(\eta_{i^*} = k^* | \eta_{-i^*}, z_{i^*} = l^* \neq m+1, z_{-i^*}; \alpha) \\ = \; & \frac{\alpha^{l^*} + n_{k^*,-i}^{l^*}}{K^{l^*} \alpha^{l^*} + \sum_k n_{k,-i}^{l^*}} \end{aligned}$$

To derive the conditional for x_{i^*}, recall that it is a vector of words drawn from a vocabulary of V unique words. If u be the index of unique words in x_{i^*} and each unique word be M_u times repeated in x_{i^*}, then when $z_i = m+1$

$$\begin{aligned} & p(x_{i^*1} = v_i^*, x_{i^*2} = v_2^*, \dots, x_{i^* n_{i^*}} = v_{n_{i^*}}^* | \eta_{-i^*}, \eta_{i^*}, x_{-i^*}; \beta) \\ & = \prod_{u=1}^{U} \prod_{m=0}^{M_u - 1} \frac{\beta + \sum_{i \neq i^*}^{N} \sum_{j=1}^{n_i} \delta_v(x_{ij}) \delta_{k^*}(\eta_i) + m}{V\beta + \sum_{i \neq i^*}^{N} n_i \delta_{k^*}(\eta_i) + \sum_{u'=1}^{u'<u} M_{u'} + m} \end{aligned}$$

Else, for $1 \leq z_i \leq m$, the each individual word x_{ij} in x_i is just transferred from the corresponding component in ϕ^{z_i}:

$$p(x_{i^*} | \eta_{i^*}, x_{-i^*}, \{\phi_k^1\}, \dots, \{\phi_k^m\}, \lambda^0) = \prod_w \phi_{k,w}^{z_i}$$

The two models differ in the conditional for z_i, and we address them separately.

In the case of the POG-DP model, z_i is a scalar random variable. The posterior for z_i is given as

$$p(z_{i^*} = m^* | z_{-i^*}; \beta) = \frac{\beta_{m^*} + \sum_{i \neq i^*} \delta_{m^*}(z_i)}{\sum_s \beta_s + N - 1} \quad (9)$$

where m^* picks up one of the existing m groups.

Mixing of the underlying Markov Chain can be improved by sampling (z_i, η_i) as a block for each data item:

$$\begin{aligned} & p(\eta_{i^*}, z_{i^*} | x, \eta_{-i^*}, z_{-i^*}, \phi; \alpha, \beta, \lambda) \propto p(z_{i^*} | z_{-i^*}; \beta) \\ & \quad p(\eta_{i^*} | \eta_{-i^*}, z_{i^*}, z_{-i^*}; \alpha) p(x_{i^*} | \eta_{i^*}, x_{-i^*}, \eta, \lambda) \quad (10) \end{aligned}$$

In the case of combinatorial model POG-CDP, recall that z_i is random binary vector, and therefore, sampling z_i as a block becomes infeasible because of the combinatorial

space of possible assignments. However, armed with the independent parameterization for each position of the vector, we can sample each position z_{ij} independently:

$$p(z_{ij} = m | z_{-i^*}; \rho) \propto \rho_m^j + \sum_{i \neq i^*} \delta_m(z_{ij}) \quad (11)$$

Arguably, the underlying Markov Chain can mix slower compared to the block sampler for POG-CDP, but on the other hand, this strength of this inference scheme is its able to handle a large number of overlapping groups.

V. Evaluation

In this section, we experimentally evaluate the performances of the two proposed models for data with partially observed groups, POG-DP and its combinatorial counter-part POG-CDP, for the task of topic discovery for vernacular news, as well as two other applications of topic analysis using observed source topics, — issue discovery in customer feedback analysis and emerging topic discovery in news group postings.

A. Experimental Set-up

We first describe the baselines for comparison and the evaluation methods.

Baselines: We compare against the following baselines, some of which have no access to the source at all, while others have different levels of information about the source — the actual data items, with or without topics, or the source topics like the proposed models.

Dirichlet Process (DP): This is the traditional DP Mixture Model [2], [3] that clusters the target document collection $\mathcal{D}_u$, without any knowledge about the source $\mathcal{D}_o$, following Equation (1).

Sequential Dirichlet Process (SeqDP): This baseline has access to the source data items x_i^o and source labels z_i^o in $\mathcal{D}_o$, but not the source topics η_i^o. Here the traditional DP Mixture Model first learns the source topics η_i^o from the source data x_i^o using Equation (1). With the unique topics ϕ_i and their counts m_i initialized from the source data, it next clusters the target data x_i^o, again using the DP mixture model (Equation (1)). Note that this model does not distinguish between different sources. If the multiple sources are present in the $\mathcal{D}_o$, it merges all of them into a single source.

Partially Observed Dirichlet Process (PO-DP): This can be imagined as a variant of SeqDP with partially observed topics. It has access to the true topics η_i^o for the source data items. We assume that a DP has already run on the source data and generated the *true topic assignment* η_i^0. So we initialize the unique topics ϕ_j and their counts m_j using the provided source topics η_i^o, and then proceed to cluster the target data as in a DP (Equation (1)). Note that this model also does not distinguish between different sources. If the multiple groups are present in the source data, it merges all of them into a single source.

Partially Observed Hierarchical Dirichlet Process (PO-HDP): This is the model described in Subsection III-C1. This baseline has access to the true source topics η_i^o and also distinguishes between the different sources.

Pair-wise Constrained DP Mixture Model (PC-DP): As our final baseline, we consider the constrained DP Mixture model [15], where pair-wise must-link and cannot-link constraints are provided over a subset of data items. This is in the same spirit as the popular pair-wise constrained k-means [16] used for semi-supervised clustering. Instead of providing source topics, we add *all* source data points corresponding to the source clusters as additional data points to the target dataset. Over pairs of these new source data items, we add must-link constraints if they have the same true cluster label in the source, and cannot-link constraints otherwise. Then we cluster this augmented target dataset using the constrained DP mixture model. Note that this baseline also cannot distinguish between different sources in $\mathcal{D}_o$.

Evaluation measures: In all three applications, we have gold standard cluster labels for the target data set, corresponding to user interpretable or meaningful topics. We evaluate the performance of all of the models quantitatively by measuring how well the discovered topics correspond to gold standard topics which the user intended to find. We measure the accuracy of pair-wise clustering decisions over all target documents using the F1-measure (Overall F1), defined as the harmonic mean of precision and recall. As a finer grained evaluation, we also separately consider the target documents that truly correspond to preferred topics in the source and those that do not. Source F1 is calculated over pairs of documents corresponding to source topics in the gold-standard, and measures how well target documents known source topics are identified. Target F1 is calculated over pairs of all other data points, and measures how well a model identifies target data points corresponding to new topics.

We first consider a single source setting, where preferred topics are provided for one source dataset. Note that the combinatorial HDP is not applicable for this setting, so we compare the POG-DP against various baselines. Then we consider the multi-source setting, where we additionally compare the POG-DP with the POG-CDP.

B. Experiments with single source

We first evaluate the proposed models and the baselines in the single-source setting for the three applications.

Vernacular News Analysis (VNA): As our first application, we consider the novel task of discovering news topics from a vernacular news collection. We consider news articles from a regional language news-paper (B)[1] from 01-2007 to 12-2007. We assume that we have the knowledge of news

[1] http://www.anandabazar.com/

topics from an English language newspaper (E)[2] published from the same city over the same period, but catering to a different population segment. Now, the task is to find out the topics in B that correspond to some English topic from E, and the novel topics that are reported exclusively in B. Observe that the asymmetry of the task arises naturally, since meaningful topics are easier to detect in English using the various linguistic and semantic resources available. In case of news in other languages, where such resources are still hard to come by, a practical approach is to identify English news topics, and detect other novel ones.

To set up this experiment, we make use of domain knowledge and decide on 10 news topics, 5 of which are of national importance and are reported in both B and E, and the remaining 5 relate to regional news reported only in B. Then we extract the news stories from B that correspond only to these 10 topics, using a seeded variant of the popular LDA model [10]. This results in a target collection $\mathcal{D}_u$ of 2000 documents over a vocabulary of 5000 words. We specify observed topics ($\mathcal{D}_o$) using a single source containing 500 news articles from the 5 common topics.

Customer Feedback Analysis (CFA): As our second application, we experiment on real data from customer service analysis. The task is to discover significant issues mentioned by customers of a Tele-communication company (which we call company Company T, for confidentiality reasons) in customer satisfaction surveys. As prior information, we have available a previously analyzed collection of surveys for a web-service provider company (Company W1), with individual feedbacks labeled with issues. Given this, we need to find out if the same issues are relevant for Company T, or if some new issues are also involved. This is a very challenging task, where the feedbacks contain free-form text with abundant spelling mistakes and abbreviations. Gold-standard issues are available for a subset of Company T's data for evaluation.

We created $\mathcal{D}_o$ having a single source containing 2 issues from Company W1 relating to call centers (*communication problems* and *timely response*). We created the target collection $\mathcal{D}_u$ using feedbacks from Company T corresponding to 5 different issues — the 2 call-center related issues, and 3 additional issues (*product*, *policy* and *web-site*). This resulted in a target collection containing 500 documents over a vocabulary of 1200 unique words.

News-Group Analysis (NGA): In the third application, we look at the task of identifying breaking discussion topics in news groups. Specifically, given the existing discussion categories known to the moderators, and the postings over a period of time, we want to cluster the postings either according to the existing categories, or according to new ones. For this task, we use the 20 Newsgroups dataset[3]. The target collection $\mathcal{D}_u$ consists of all postings from all 20 available categories ($\sim 14,000$ postings using ~ 5000 unique words). The source collection $\mathcal{D}_o$ is created using all 6 categories under $comp$ and $misc$ (~ 5000 postings).

F1	DP	Seq DP	PO-DP	PC-DP	PO-HDP	POG-DP
VNA	0.25	0.51	0.63	0.59	0.59	**0.71**
CFA	0.26	0.26	0.33	0.31	0.32	**0.39**
NGA	0.34	0.34	0.44	0.44	0.45	**0.53**

Table I
OVERALL F1 FOR VERNACULAR NEWS ANALYSIS, CUSTOMER FEEDBACK ANALYSIS AND NEWS-GROUP ANALYSIS (500 SAMPLES)

	DP	Seq DP	PO-DP	PC-DP	PO-HDP	POG-DP
Source	0.26	0.44	0.72	0.64	0.80	**0.88**
Target	0.28	0.26	0.31	0.29	0.30	**0.34**

Table II
PERFORMANCE FOR SOURCE AND TARGET ISSUES IN CUSTOMER FEEDBACK ANALYSIS

Results for Single Source: The Overall F1 in the three tasks for the various models are reported in Table I. In summary, the POG-DP model performs the best overall across all three tasks. The three partially observed baselines (PC-DP, PO-DP and PO-HDP) expectedly perform better than the other two. However, POG-DP is able to outperform them essentially through the coupling of the component selection probabilities within and outside the source, and also by allowing topic preferences to be learnt for the sources. Table II records the Source F1 and Target F1 scores separately for Customer Feedback Analysis. This shows that POG-DP does well on both aspects of the task, recognizing known issues and identifying new ones. The same trend appears for the other two tasks as well.

[mrinal]We have searched over a small interval over the hyper-parameters and the one giving the best result has been reported here.[/mrinal] The transfer hyper-parameter values (β_0, β_1) for POG-DP used in the experiments are $(250, 500)$ for VNA, $(100, 100)$ for CFA and $(1000, 1000)$ for NGA. The other topic hyper-parameters $(\alpha_0 = 0.01, \alpha_1 = 0.01, \lambda = 1)$ are the same for all models.

C. Experiments with Multiple Sources

It is natural in all of these three applications, to have known topics specified through multiple sources simultaneously. Next, we evaluate the performance of POG-DP and its combinatorial version POG-CDP in such settings.

First, for vernacular new analysis, we can have available the news categories from multiple national newspapers, and then look for regional news topics from local language news

[2]http://www.telegraphindia.com/section/frontpage/index.jsp

[3]http://people.csail.mit.edu/jrennie/20Newsgroups

	PO-DP	PC-DP	PO-HDP	msPOG-DP	POG-DP	POG-CDP
VNA	0.71	0.69	0.73	0.81	0.81	**0.86**
CFA	0.33	0.31	0.34	0.37	0.325	**0.43**
NGA	0.60	0.60	0.63	0.67	0.62	**0.69**

Table III
OVERALL F1 WITH MULTIPLE OVERLAPPING SOURCES

sources. To set-up the experiment, we consider news articles from a second vernacular language newspaper[4], which we call H, again over the same time-period, but catering to a third population segment, possibly overlapping with the other two. To create the first source for $\mathcal{D}_o$, we identify 7 topics from E, and for the second source, 7 from H, such that 4 of these overlap, and all 10 appear in B. To create the target collection $\mathcal{D}_u$, in addition to the 10 topics shared with E and H, we include news articles from 5 topics exclusive to B. This results in a target collections of 3000 news stories from 15 topics.

We provide $\mathcal{D}_o$ and $\mathcal{D}_o$ as input to POG-DP and POG-CDP. Recall that the two sources in $\mathcal{D}_o$ have 4 shared news topics. Therefore, POG-CDP creates 3 source intersections, the first two containing 3 topics exclusive to E and H respectively, and the third containing the 4 topics that appear in both. Note that some of the baselines (SeqDP, PO-DP, PC-DP) take the union of all the topics from the multiple sources and create a single source with those topics. So we evaluate the performance of a merged source version of POG-DP as well (msPOG-DP).

We set up similar experiments for customer service analysis and news-group analysis as well. For customer service analysis, we create a second source using customer feedback data from a second web-service providing company pertaining to two issues, *communication problem* and *website*, both of which are relevant for the target data from Company T, and one is shared with the first source from Company W1. For news-group analysis, we construct two sources (S1 = [talk.politics.misc & talk.religion.misc], S2 = [soc.religion.christian, alt.atheism & talk.religion.misc]) with one topic (talk.religion.misc) in common. The target collection consists of the 4 topics from the two sources, and two additional topics each from comp and rec.

The experimental results with multiple sources are reported in Table III. We only report the performance of the supervised baselines, PO-DP, PO-HDP and PC-DP. We see our combinatorial model (POG-CDP) comes out as the best consistently in all three tasks, highlighting the need to handle overlapping subsets of sources. The POG-DP and its merged source version msPOG-DP are also able to outperform the DP-based baselines PO-DP and PC-DP, which do not distinguish between different sources. The PO-HDP captures the knowledge of different sources, but the msPOG-DP model outperforms it by virtue of (1) modeling source-specific preferences that couple topic selection probabilities within a source, and (2) not modeling the source topics. However, the combinatorial model performs even better by coupling finer granularities of source topics. POG-DP does not perform significantly better than merged-source version msPOG-DP, largely on account of the overlap between sources. In other experiments, where the sources are mutually exclusive in terms of issues, POG-DP outperforms msPOG-DP.

Summary & Discussion: The experiments demonstrate that the Dirichlet Process models with partially observed groups are able to effectively identify topics from source collections as well as discover new topics that appear exclusively in the target. In all three applications, the POG-DP model outperforms partially observed variants of DP and HDP, as well as pair-wise constrained DP, by virtue of avoiding generative assumptions on source topics and coupling topic selection probabilities within a source. In situations where multiple overlapping sources are available, the combinatorial model POG-CDP performs the best. It is also able to identify coherent subsets of topics from different sources that are relevant for the target. [mrinal] Moreover, the computational cost is very less (only few seconds).[/mrinal]

VI. CONCLUSIONS AND FUTURE WORK

Motivated by the task of vernacular news analysis, we study the problem of topic analysis in a target dataset, given sources with observed topics. We have proposed Dirichlet Process with partially observed groups, that directly models the conditional distribution of the target data conditioned on the source topics using coupled Dirichlet Processes. This improves over the HDP, that unnecessarily and often inappropriately models the generative process for observed topics. The POG-DP model also introduces coupling between source topics, and are able to identify them more effectively. We improve over this further for overlapping sources, with the combinatorial Dirichlet Process model that parameterizes arbitrary subsets of the sources to introduce fine-grained coupling between selection probabilities of source topics. We have proposed efficient inference algorithms for these models. We have demonstrated the usefulness of the proposed models through extensive experiments over various baselines for three different real-life applications.

REFERENCES

[1] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *Jour. of ASA*, 2006.

[2] T. Ferguson, "A bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, 1973.

[3] C. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *Ann. Statist.*, vol. 2, no. 6, pp. 1152–1174, 1974.

[4]http://in.jagran.yahoo.com/epaper/

[4] P. Muller, F. Quintana, and G. Rosner, "A method for combining inference across related nonparametric bayesian models," *Jour. of the Royal Stat. Society*, vol. 66, pp. 735–749, 2004.

[5] S. MacEachern, "Dependent nonparametric processes," *ASA Proc. on Bayesian Statistical Science*, pp. 50–55, 1999.

[6] D. Lin, E. Grimson, and J. Fisher, "Construction of dependent dirichlet processes based on poisson processes," in *NIPS*, 2010.

[7] A. Rodrguez, D. Dunson, and A. Gelfand, "The nested dirichlet process," *Jour. of ASA*, vol. 103, no. 483, 2008.

[8] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "An hdp-hmm for systems with state persistence," in *ICML '08*, 2008.

[9] Z. Shen, P. Luo, S. Yang, and X. Shen, "Topic modeling ensembles," in *ICDM*, 2010, pp. 1031–1036.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[11] D. Blackwell and J. MacQueen, "Ferguson distributions via polya urn schemes," *Ann. of Stats.*, vol. 1, pp. 353–355, 1973.

[12] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[13] R. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[14] J. Ishwaran and L. James, "Gibbs sampling methods for stick-breaking priors," *Jour. of ASA*, vol. 96, pp. 161–174, 2001.

[15] A. Vlachos, A. Korhonen, and Z. Ghahramani, "Unsupervised and constrained dirichlet process mixture models for verb clustering," in *Workshop on Geometrical Models of Natural Language Semantics, ACL*, 2009.

[16] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML '01*, 2001.