The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Topic Modeling on Health Journals with Regularized Variational Inference

Robert Giaquinto, Arindam Banerjee
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities
{smit7982@umn.edu, banerjee@cs.umn.edu}

Abstract

Topic modeling enables exploration and compact representation of a corpus. The CaringBridge (CB) dataset is a massive collection of journals written by patients and caregivers during a health crisis. Topic modeling on the CB dataset, however, is challenging due to the asynchronous nature of multiple authors writing about their health journeys. To overcome this challenge we introduce the Dynamic Author-Persona topic model (DAP), a probabilistic graphical model designed for temporal corpora with multiple authors. The novelty of the DAP model lies in its representation of authors by a persona — where personas capture the propensity to write about certain topics over time. Further, we present a regularized variational inference (RVI) algorithm, which we use to encourage the DAP model's personas to be distinct. Our results show significant improvements over competing topic models — particularly after regularization, and highlight the DAP model's unique ability to capture common journeys shared by different authors.

1 Introduction

Topic models can compactly represent large collections of documents by the themes running through them. We introduce a topic model designed for the unique challenges presented by the CaringBridge (CB) dataset. The CB dataset includes journals written by patients and caregivers during a health crisis. CB journals function like a blog, and are shared to a private community of friends and family. The full dataset includes 13.1 million journals written by approximately half a million authors between 2006 and 2016. From the CB dataset we're interested in capturing health journeys, that is, authors writing about the same topics over time.

The challenges in topic modeling on the CB dataset stem from the asynchronous nature of author's posts. Specifically, authors start and stop journaling at different times — both in terms of calendar dates and how far along they are in their health journey. Additionally, authors post at irregular frequencies. While about 15% of CB authors post nearly everyday, the majority of authors typically post less frequently, often corresponding to a major update, event, or anniversary of an event. What's more, the length of these posts can range from just a few words to thousands of words.

State-of-the-art topic models can identify topics (Blei, Ng, and Jordan 2003), track how topics change over time (Blei and Lafferty 2006; Wang and McCallum 2006; Wei, Sun, and Wang 2007; Wang, Blei, and Heckerman 2008), or associate authors with certain topics (Rosen-Zvi et al. 2004; Steyvers et al. 2004; McCallum, Corrada-Emmanuel, and Wang 2005; Mimno and McCallum 2007). These models cannot, however, describe common narratives and the authors sharing them. We present the Dynamic Author-Persona topic model (DAP), a novel approach that represents authors by latent *personas*. Personas act as a soft-clustering on authors based their propensity to write about similar topics over time. Our approach is unique in multiple respects. First, unlike other temporal topic models, the words making up a topic don't evolve over time — rather, DAP's personas reflect the flow of conversation from one topic to next. Second, we introduce a regularized variational inference (RVI) algorithm, an approach we use to encourage personas to be distinct from one another.

Our results show that the DAP model outperforms competing topic models, producing better likelihoods on held-out data. Finally, we demonstrate that using RVI further improves the DAP model's performance, and results in personas that are rich and compelling descriptions of the health journeys experienced by CB authors.

The rest of the paper is as follows: Section 2 is a brief background on temporal topic models. Section 3 presents the DAP model. Section 4 details the model's RVI algorithm. Section 5 introduces the evaluation dataset and procedure. Section 6 shares the results of the experiments. Finally, in Section 7 we summarize the contributions of this paper.

2 Background

Much of the research on topic modeling builds on the latent Dirichlet allocation (LDA) model (Blei, Ng, and Jordan 2003). The LDA model doesn't account for meta-information like authorship or time. Nevertheless, interest in LDA has endured, in part, due its ability to richly describe topics as distributions over words and documents as mixtures of topics. In the years since LDA's introduction, others have extended the idea to compliment corpora with a variety of structures and metadata.

Author information is common in many corpora. A few topic models have been designed to identify authors' pref-

erences for certain topics, and the relationships between authors (Rosen-Zvi et al. 2004; Steyvers et al. 2004; McCallum, Corrada-Emmanuel, and Wang 2005; Mimno and McCallum 2007; Pathak et al. 2008). Corpora with a temporal structure, such as scientific journals or newspaper articles, are the focus of a number of temporal topic models (Blei and Lafferty 2006; Wang and McCallum 2006; Wei, Sun, and Wang 2007; Wang, Blei, and Heckerman 2008).

Temporal Topic Models. Two topic models set the standard of comparison for topic modeling on corpora with a temporal element: the dynamic topic model (DTM) (Blei and Lafferty 2006) and the topics over time model (TOT) (Wang and McCallum 2006). These two models represent very different approaches to modeling time in a topic model.

The TOT model defines time as an observed variable, which leads to a continuous treatment of time and the ability to predict timestamps of documents. Alternatively, the DTM evolves topics over time using a Markov process. In many corpora the evolution of topics provides interesting insights. For example, Blei's model of the *Science* corpus shows words associated with a topic on physics changing over a century.

Building directly on the DTM, in 2008 Wang et al. developed the continuous time dynamic topic model (CDTM) which uses continuous Brownian motion to model the evolution of topics over time (Wang, Blei, and Heckerman 2008). This is a major development in temporal topic models because, unlike the DTM, it doesn't require partitioning the data into discrete time periods. Instead, the model assumes that at each time step the variance in the topic proportions increases proportional to the duration since the previous document. Similar to Wang et al., the Dynamic Mixture Model (DMM) is built for continuous streams of text (Wei, Sun, and Wang 2007). In the DMM, however, topics are fixed in time and the model captures the evolution of document-level topic proportions over time.

Topic Modeling of Health Journeys. In many topic modeling applications to temporal corpora, the time component is ignored. For example, Wen et al. model cancer event trajectories from users of an online forum for breast cancer support (Wen and Rose 2012). Wen's approach uses LDA to extract cancer event keywords, which are then linked together in time by temporal descriptions mined from the text. This work demonstrates a quantitative approach to studying the dynamics of social support network, and offer a powerful look at the experiences of users in these support networks.

Numerous studies have shown that support networks, both in person and online, are valuable tools for those suffering from chronic conditions or life-threating illness and caregivers (Wen et al. 2011; Rodgers and Chen 2005; Beaudoin and Tao 2007). Additionally, online social networks can serve as a way to efficiently disseminate information regarding someone's status to their community. Understanding the health journeys of users in these social support communities is valuable information for improving user experience. Topic models are uniquely suited to succinctly describing and analyzing these health journeys.

3 The DAP Model

The design of the DAP model was made with journaling behavior in mind. Consider a CB author journaling about their surgery: initially they may write about topics related to the surgical procedure, but as time progresses the author is more likely to discuss recovery, physical therapy, or returning to normal life. In other words, the likelihood of a topic for some document depends where the document's author is in their health journey. As such, DAP assumes that (1) a state space model controls the likelihood of a topic at each time step, (2) each persona represents a different flow of topics over time, and (3) each author has a distribution over personas.

The DAP's approach for modeling topics in a document, and words in a topic follows the correlated topic model (CTM) and LDA, respectively (Lafferty and Blei 2006; Blei, Ng, and Jordan 2003). The idea of modeling latent personas was originally proposed by Mimno and McCallum, however in their Author-Persona Topic model (APT) personas differ significantly from those proposed in the DAP model. First, DAP models each author as a distribution over a fixed number of personas. Second, we model documents as each having their own distribution over topics. Lastly, while DAP's personas also correspond to a distribution over topics, DAP evolves these topic distributions over time — thereby capturing the inherent temporal structure resulting from an author writing multiple documents.

The DAP model directly addresses the challenges presented by the CB dataset. First, the asynchronous nature of health journals is handled by: (1) transforming each journal's timestamp to the time elapsed since the author's first post, and (2) learning multiple personas to account for a wide variety in topic trajectories. Second, irregular posting behavior is managed by employing the Brownian motion model, originally used in topic modeling by Wang, Blei, and Heckerman to model topic variance as proportional to the gap in time between documents.

The generative process of the model is described below. The model assumes that each document d in the corpus has a timestamp s_t associated with it. Similar to the CDTM (Wang, Blei, and Heckerman 2008), timestamps are used in a continuous Brownian motion model to capture an increase in topic variance as time between observations increases. More formally, if s_i and s_j are timestamps at steps $j > i > 0$, then Δ_{s_j,s_i} is the difference in time between s_j and s_i. We use the shorthand Δ_{s_t} to denote the difference in time between timestamps s_t and s_{t-1}. For brevity the variance $\sigma\Delta_{s_t}I$ is denoted Σ_t, where σ is a known process noise in the state space model.

1. Draw distribution over words $\beta_k \sim Dir(\eta)$ for each topic k.
2. Draw distribution over personas $\kappa_a \sim Dir(\omega)$ for each author a.
3. For each persona p, draw initial distribution over topics:

$$\alpha_{0,p} \sim \mathcal{N}(\mu_0, \Sigma_0), \forall p \in \{1, \ldots, P\} .$$

4. For each time step t, where $t \in \{1, \ldots, T\}$:

- Draw distribution over topics:

$$\alpha_{t,p} \sim \mathcal{N}(\alpha_{t-1,p}, \Sigma_{t-1}), \forall p \in \{1, \ldots, P\} .$$

- Update Σ_t according to Brownian motion model: $\Sigma_t - \Sigma_{t-1} \sim \mathcal{N}(0, \sigma\Delta_{s_t} I)$.
- For each document d, where $d \in \{1, \ldots, D_t\}$:

(a) Choose persona indicator $x_{t,d} \sim Mult(\kappa_a)$ where a corresponds to the author of document d_t.
(b) Draw topic distribution $\theta_{t,d} \sim \mathcal{N}(\alpha_t x_{t,d}, \Sigma_t)$ for document d_t.
(c) For each word $w_{t,d,n}$, where $n \in \{1, \ldots, N_{d_t}\}$:
  i. Choose word topic indicator $z_n \sim Mult(\pi(\theta_{t,d}))$.
  ii. Choose word $w_{t,d,n}$ from $p(w_{t,d,n} \mid \beta_{z_n})$, a multinomial probability conditioned on the topic indicator z_n.

Following the approach in the CTM and DTM, we use the function $\pi(\cdot)$ to map the Logistic Normal $\theta_{t,d}$, parameterized by a mean $\alpha_{t,k,p}$ and covariance $\sigma\Delta_{s_t} I$, to the multinomial's natural parameters via $\pi(\theta_{t,d}) = \frac{\exp(\theta_{t,d})}{\sum_d^D \exp(\theta_{t,d})}$ in order to obey the constraint that the parameters lie on the simplex.

The graphical model corresponding to this process is shown in Figure 1. In LDA and its extensions the parameter α represents a prior probability of each topic. In the DAP model, $\alpha_{t,1:K,p}$ takes on an expanded role: it's a distribution over K topics at time step t for persona p. The choice of letting α evolve over time, as opposed to β like in the DTM, is that in a collection of journals there is less interest in changes to topics themselves. In other words, we model the words associated with a topic as static in time, but the topics an author writes about will change over time.

4 Variational EM Algorithm

Given the model structure, next we derive an inference algorithm used to estimate the model's latent parameters. Much like LDA and its extensions, the DAP model's posterior:

$$p(\kappa, \mathbf{x}, \alpha, \beta, \theta, \mathbf{z} \mid \mathbf{w}, \omega, \eta) = \frac{p(\kappa, \mathbf{x}, \alpha, \beta, \theta, \mathbf{z}, \mathbf{w} \mid \omega, \eta)}{p(\mathbf{w} \mid \mu_0, \sigma_0, \eta, \omega)} ,$$

is intractable due to the normalization term. In order learn optimal values to the model's parameters we use a form of variational inference (VI), which approximates the difficult to compute posterior distribution p with a simpler distribution q (see Blei, Kucukelbir, and McAuliffe, 2016 for a review). Variational inference casts an inference problem as an optimization problem with the goal of finding parameters to the variational distribution such that $q = q(\kappa, \mathbf{x}, \alpha, \theta, \mathbf{z}, \beta)$ closely approximates $p = p(\kappa, \mathbf{x}, \alpha, \theta, \mathbf{z}, \beta \mid \mathbf{w})$. Our regularized variational inference (RVI) algorithm seeks a distribution $q \in \mathcal{Q}$ such that

$$q^* = \arg\min_{q \in \mathcal{Q}} KL(q \,||\, p) + \rho r(q) , \tag{1}$$

where $KL(\cdot)$ is KL-Divergence. The added term $r(q)$ is a regularization function we've introduced to discourage similar personas (further detail given in Section 4.2), and ρ the corresponding hyperparameter.

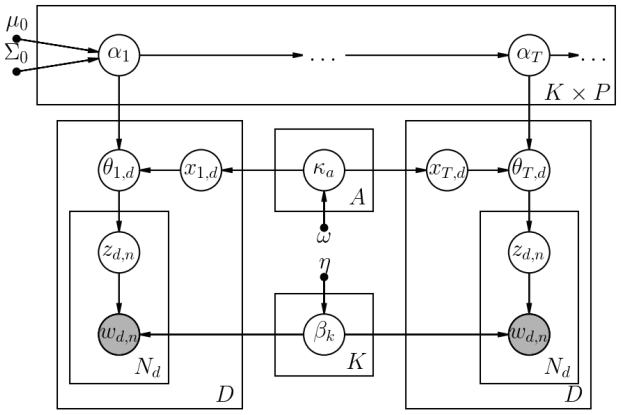

Figure 1: Graphical representation of the Dynamic Author-Persona topic model (DAP). On top, topic distributions for each persona evolve over time: $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \Sigma)$. The distribution over words for each topic, $\beta \sim Dir(\eta)$, is fixed in time. Each author $a \in \{1, \ldots, A\}$ is represented by a distribution over personas, that is $\kappa_a \sim Dir(\omega)$. The distribution over topics for each document is dependent on the persona distribution $\mathbf{x}_{t,d}$ for that document's author, and the evolving topic distribution α_t.

To make q easy to compute, we apply mean field variational inference which assumes that the parameters are *posteriori* independent. Under the mean field assumption the variational distribution factorizes as:

$$\prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{a=1}^{A} q(\kappa_a \mid \delta_a) \prod_{p=1}^{P} q(\alpha_{1:T,k,p} \mid \hat{\alpha}_{1:T,k,p}) \times \prod_{t=1}^{T} \prod_{d=1}^{D_t} q(x_{t,d,p} \mid \tau_{t,d,p}) q(\theta_{t,d} \mid \gamma_{t,d}) \prod_{n=1}^{N_{d_t}} q(z_n \mid \phi_n) \tag{2}$$

where we have introduced the following variational parameters: the persona for each author κ_a is endowed with a free Dirichlet parameter δ_a; each assignment of a persona to an author $\mathbf{x}_{t,d}$ is endowed with a free Multinomial parameter $\tau_{t,d}$; in the variational distribution of $\alpha_{1:T,k,p}$ the sequential structure is kept intact with variational observations $\hat{\alpha}_{1:T,k,p}$; each document-topic proportion vector $\theta_{t,d}$ is endowed with a free γ_d. The variance for the document-topic parameters are $v_{t,d}$ and $\hat{v}_{t,d}$, for the model and variational parameter, respectively; each word-topic indicator is endowed with a free multinomial parameter $\phi_{t,d,n}$.

Optimization of the variational parameters proceeds using variational expectation-maximization. The objective function in (1) cannot be computed directly, we therefore maximize a surrogate likelihood consisting of the Evidence Lower BOund (ELBO) minus the regularization term (see Wainwright and Jordan, 2007 for a review):

$$\mathcal{L}_\rho(\delta_a, \tau_{t,d}, \gamma_{t,d}, \phi_n, \lambda_k) \triangleq \mathbf{E}_q[\log p] - \mathbf{E}_q[\log q] - \rho r(q) \tag{3}$$

Expanding the objective function $\mathcal{L}_\rho$ according to the distribution associated with each parameter allows updates to

be derived for each parameter. The parameters are optimized using a variational expectation-maximization algorithm, the details of the algorithm are given below.

4.1 Variational E-Step

During the E-step the model estimates variational parameters for each document and saves the sufficient statistics required to compute global parameters. The structure of the DAP model, while unique, has some components that mimic previous topic models. Specifically, the word-topic assignment parameter ϕ has the same update found in the CTM due to the Logistic-Normal γ parameter. Hence ϕ has a closed form update: $\phi_{n,k} \propto \exp(\gamma_k)\beta_{k,v}$ (Lafferty and Blei 2006).

Each author's persona is parameterized by a τ. To find an update for τ we select ELBO terms featuring τ, along with the Lagrangian term λ to ensure each vector $\tau_{t,d}$ sums to one, and then take the derivative with respect to each document and persona.

$$\frac{\partial \mathcal{L}}{\partial \tau_{t,d,p}} = \Psi(\delta_{a,p}) - \Psi(\sum_{i=1}^{P} \delta_{a,i}) - \log \tau_{a,p} - 1 + \lambda + \hat{\alpha}_{t,p}\Sigma_t^{-1}(\gamma_{t,d} - \hat{\alpha}_{t,p}\tau_{t,d,p}) - \frac{1}{2}\operatorname{Tr}(\Sigma_t^{-1}\operatorname{diag}(\hat{\alpha}_{t,p}^2 + \hat{\Sigma}_t))$$

Since a closed form solution for $\tau_{t,d}$ doesn't exist, we therefore estimate $\tau_{t,d}$ using exponential gradient descent.

Since the model includes non-conjugate terms, an additional variational parameter ζ is introduced to preserve the lower-bound during the expansion of the non-conjugate pairs term $\mathbf{E}_q[\log p(\mathbf{z}_n \mid \pi(\theta_{t,d}))]$. Taking the derivative of all terms containing ζ and setting it to zero yields an analogous closed form update to the one found in the CTM: $\hat{\zeta}_t = \sum_{k=1}^{K} \exp(\gamma_{t,d,k} + \hat{v}_{t,k}^2/2)$.

Finally, the DAP model estimates a topic distribution for each document via the $\gamma_{t,d}$ parameter. A conjugate gradient algorithm is run using the gradient:

$$\frac{\partial \mathcal{L}}{\partial \gamma_{t,d,k}} = -\Sigma_t^{-1}(\gamma_{t,d,k} - \hat{\alpha}_{t,1:P,k}\tau_{t,d,k}) + \sum_{n=1}^{N_{d_t}} \phi_{n,k} - \frac{N_{d_t}}{\zeta}\exp(\gamma_{t,d,k} + \hat{v}_{t,k}^2/2))$$

Whereas $\gamma_{t,d}$ represents the mean of the Logistic-Normal for a document's topic distribution, the parameter $\hat{v}_{t,d}$ is the variance. Setting the derivative of $\mathcal{L}(\hat{v}_{t,d})$ with respect to $\hat{v}_{t,d}$ to zero and solving yields:

$$\frac{\partial \mathcal{L}}{\partial v_{t,d,k}^2} = \Sigma_{t,k,k}^{-1} + \frac{1}{2\hat{v}_{t,d,k}^2} - \frac{N_{d_t}}{2\zeta}\exp(\gamma_{t,d,k} + \hat{v}_{t,k}^2/2),$$

which requires Newton's method for each coordinate, constrained such that $\hat{v}_{t,k} > 0, \forall k$.

The parameter $\hat{\alpha}_t$ represents the noisy estimate of α_t. After calculating $\hat{\alpha}_t$, the forward and backward equations will be applied in the M-step to give a final posterior estimate α_t. The terms in the ELBO containing $\hat{\alpha}_t$ are the result of expanding $\mathbf{E}_q[\log p(\alpha_{t,p} \mid \alpha_{t-1,p})]$ for (4a) and $\mathbf{E}_q[\log p(\theta_{t,d} \mid \alpha_t \mathbf{x}_{t,d}, \Sigma_t)]$ for (4b) and (4c):

$$\mathcal{L}(\hat{\alpha}) = \sum_{t=1}^{T}\sum_{p=1}^{P} -\frac{1}{2}(\hat{\alpha}_{t,p} - \hat{\alpha}_{t-1,p})^\top \Sigma_t^{-1}(\hat{\alpha}_{t,p} - \hat{\alpha}_{t-1,p}) + \tag{4a}$$

$$\sum_{t=1}^{T}\sum_{d=1}^{D_t} -\frac{1}{2}\Big((\gamma_{t,d} - \hat{\alpha}_t\tau_{t,d})^\top \Sigma_t^{-1}(\gamma_{t,d} - \hat{\alpha}_t\tau_{t,d}) + \tag{4b}$$

$$\sum_{t=1}^{T}\sum_{d=1}^{D_t}\sum_{p=1}^{P} \operatorname{Tr}\left[\Sigma_t^{-1}\operatorname{diag}\left(\tau_{t,d,p}(\hat{\alpha}_{t,p}\hat{\alpha}_{t,p}^\top + \hat{\Sigma}_t)\right)\right]\Big) \tag{4c}$$

Taking the derivative with respect to the mean term for each persona gives the closed form update:

$$\hat{\alpha}_{t,p} = \frac{\hat{\alpha}_{t-1,p} + \sum_{d=1}^{D_t}(\gamma_{t,d} + 1)\tau_{t,d,p}}{1 + \sum_{d=1}^{D_t}\tau_{t,d,p}^2} \tag{5}$$

We solve for $\hat{\alpha}_{t,p}$ sequentially over time steps. For the initial time step $t = 1$, we use the prior μ_0 in place of $\hat{\alpha}_{t-1,p}$. Note that the summations in (5) are collected during the E-step and $\hat{\alpha}_{t,p}$ need only be computed once after performing inference on all documents.

4.2 Regularized Variation Inference

Our RVI algorithm nudges α_t to find topic distributions that are different for each persona. A natural choice for capturing this idea is an inner product between each of the personas (excluding a persona with itself). Hence, we define the regularization function by:

$$\rho r(q) = \sum_{p=1}^{P}\sum_{1\le q\le P, q\neq p} \frac{D_t}{2}\rho\hat{\alpha}_{t,p}^\top\Sigma_t^{-1}\hat{\alpha}_{t,q}\ , \tag{6}$$

The parameter Σ_t^{-1} in included in the regularization for two reasons. First, it simplifies the update to $\hat{\alpha}_{t,p}$. In (4) the term Σ^{-1} appears in every term, which allows it to be factored out and canceled. By including Σ^{-1} in the regularization the same cancellation can occur. Second, since $\Sigma_t^{-1} \propto I$ then its inclusion has the effect of encouraging personas to be orthogonal to one another. We include the number of documents D_t at time t in $r(q)$ so that the regularization is applied evenly, regardless of dataset size or a skewed distribution of documents over time. After including the regularization term in (6) with the ELBO terms in (4), the regularized $\hat{\alpha}_{t,p}$ update is:

$$(1+\sum_{d=1}^{D_t}\tau_{d,p}^2)\hat{\alpha}_{t,p} + \rho D_t\sum_{q\neq p}\hat{\alpha}_{t,q} = \hat{\alpha}_{t-1,p} + \sum_{d=1}^{D_t}(\gamma_d + 1)\tau_{d,p} \tag{7}$$

Since the vector $\sum_{d=1}^{D_t}(\gamma_d + 1)\tau_{d,p}$ (of length K) is computed during the E-step, then the RHS is known. Similarly, the term $(1 + \sum_{d=1}^{D_t}\tau_{d,p}^2)$ is known, and in combination with ρD_t form the weights over the unknown vector $\hat{\alpha}_{t,p}$, also of length K. Therefore, (7) can be solved as a system of linear equations. Through experiments we've found an optimal value of $\rho \in [0, 0.5]$. The model exhibits sensitivity to the

hyperparameter ρ, if ρ is large (e.g. > 1.0) then model quality drops due to personas overfitting to a single topic. Since $\hat{\alpha}$ is only used to estimate the global parameter α during the M-step, computing $\hat{\alpha}$ isn't necessary for inference on holdout datasets.

4.3 M-Step

In the M-step the global parameters α, κ, and β are updated such that the lower bound of the log likelihood of the data is maximized. Note, the update for β is exactly the same as derived for the LDA model, and hence omitted.

The parameter δ represents the distribution over personas for each author. The closed form update for $\delta_{a,p}$:

$$\delta_{a,p} \propto \omega_p + \sum_{t=1}^{T} \sum_{d=1}^{D_t} \tau_{t,d,p} \ ,$$

shows that δ's closed form update is an average of the persona assignments, smoothed by the author-persona prior ω.

Once the variational observations $\hat{\alpha}_{t,p}$ are computed, our approach follows the variational Kalman filtering method from Wang's Continuous Time Dynamic Topic Model, see Wang, Blei, and Heckerman for further details. Specifically, we employ the Brownian motion to model time dynamics. However, because the DAP model's time-varying parameter is a distribution over latent topics, it performs best on data discretized in time (resulting in a smaller T). The forward equations mimic a Kalman filter:

$$m_{t,p} = \frac{\hat{\alpha}_{t,p} P_{t,p} + m_{t-1,p} \hat{w}_t}{P_{t,p} + \hat{w}_t}$$
$$V_{t,p} = \hat{w}_t \frac{P_{t,p}}{P_{t,p} + \hat{w}_t}$$

where $\hat{w}_t$ is the known process noise, and $P_{t,p} = V_{t-1,p} + \sigma \Delta_{s_t}$ captures the increase in variance as time between data points grows. Finally, the backward equations:

$$\alpha_{t-1,p} = m_{t-1,p} \frac{\sigma \Delta_{s_t}}{P_{t,p}} + \alpha_{t,p} \frac{V_{t-1,p}}{P_{t,p}}$$
$$\Sigma_{t-1,p} = V_{t-1,p} + \frac{(V_{t-1,p})^2}{(P_{t,p})^2} (\Sigma_{t,p} - P_{t,p}) \ ,$$

give the updates to the remaining global parameters.

5 Experiments

5.1 CaringBridge Dataset

The creation of our model is inspired by a desire to discover topics on a unique dataset consisting of 14 million journals posted by half a million authors on the social networking site CaringBridge (CB). Established in 1997, CaringBridge is a 501(c)(3) non-profit organization focused on connecting people and reducing the feelings of isolation that are often associated with a patient's health journey. Due to their content, CB data has been anonymized prior to analysis.

From the CB dataset we draw an evaluation dataset consisting of journals written by authors who posted, on average, at least twice a month over a one year period. Journal posts are only kept if they contain 10 or more words. These constraints help identify a set of active users. From the 123K authors meeting these criteria, 2,000 were randomly selected. Journals written by these 2,000 authors total 114,532. Overall, authors in this dataset journal an average of 57 times, with a mean of 5 days between journal posts.

5.2 Evaluation

Journals are split into training and test sets with 90% of each author's journals ($N = 103,018$) for training and 10% ($N = 11,728$) for testing. Further, variance in model performance is estimated by repeating this splitting procedure for 10-fold cross validation.

The performance of our model is compared to three other models representing the state-of-the-art in this area. The first model for comparison is LDA, which ignores authorship and temporal structure in the data. In order to evaluate LDA's performance over time, we train LDA on time steps up through $t-1$ and testing on time step t (similar to the evaluation method in Wang and McCallum, 2006). The DTM also serves as an important baseline for comparison because it models the evolution of topics over discrete time steps. Lastly, we compare out model to CDTM, which builds on DTM and introduces a continuous treatment of time. Following the approach of others, we simply fix the number of topics at 25 for all models. The number of personas learned by the DAP model is fixed at 15.

To evaluate the models we compute the per-word log-likelihood ($PWLL$) on heldout data, which measures how well the model fits the data and is computed by $PWLL = \frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}$. Note that perplexity, another common metric used to compare topic models, is related to $PWLL$ via $perplexity = \exp(-PWLL)$. It has been shown that perplexity (and hence $PWLL$s) don't correlate with a model finding coherent topics (Chang et al. 2009). Nevertheless, $PWLL$s provide a fair way to compare how well each model optimizes their objective functions.

6 Results

In addition evaluating model fit, we perform a qualitative analysis of the DAP model to highlight the quality and usefulness of the personas discovered. In particular, we establish that the personas are unique from one-another and capture meaningful experiences shared by authors.

6.1 Model Comparison

In Table 1 we list the per-word log-likelihood and standard deviation between cross-validation sets for each of the competing models. There is a significant improvement in the DAP model's performance after regularization. Further analysis of the likelihood computation reveals that the regularization term contributes a relatively small drop in likelihood compared to the total likelihood during training. Nevertheless, these results show that even a small amount of regularization can nudge the model to seek out quality results. In testing additional ρ values we found that, in general, $\rho \in [0.1, 0.3]$ faired comparably. Larger values of ρ can cause model instability and the document likelihoods

Model	Per-word Log-Likelihood	Std. Dev.
DAP (ρ=0.0)	-7.22	0.04
DAP (ρ=0.2)	-6.47	0.04
LDA	-9.23	0.02
DTM	-9.65	0.03
CDTM	-8.82	0.03

Table 1: Overall comparison of models. Per-word log-likelihoods for documents in the test dataset are computed. Standard deviation in performance computed over the cross-validation sets. While the basic DAP model without regularization performs significantly better than competing model, the RVI approach further increases log-likelihoods.

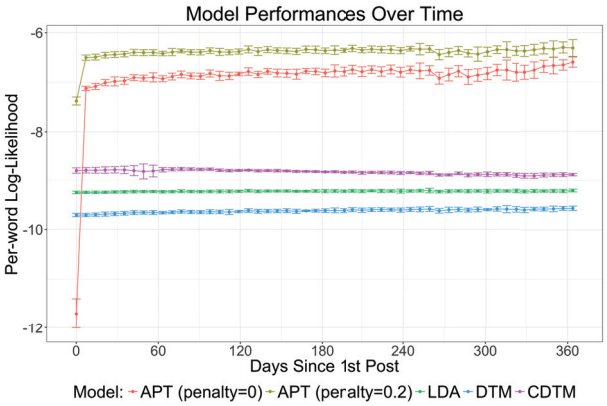

Figure 2: In general the DAP model performs better than competing models over time steps. The regularized DAP model further improves performance and reduces variable results found in the first time step of the unregularized model. Error bars show one standard deviation in document-level PWLL.

to have long-tailed distributions. The emergence of outlier document-likelihoods is unsurprising, regularization encourages the personas to focus on different topics — hence, large values of ρ inevitably result in personas that overfit.

Figure 2 shows mean per-word log-likelihoods at each time step. The best performing DAP model shows consistently better results over competing models. However, the unregularized DAP model has a significant drop in performance in the first time step.

6.2 Persona Quality

To evaluate personas we focus on three key elements: authors are described by one persona, personas are distinct, and personas capture coherent health journeys.

1:1 Author-Persona Mappings. Authors are modeled as a distribution over personas; however, to create interpretable results we want these distributions to focus on a single persona. The DAP model achieves this in the majority of cases: 71% of authors are concentrated on a single persona ($> 90\%$ probability for that persona), and 27% of authors are evenly split between two personas. This shows that, in general, the model finds personas that generalize well enough to describe the majority of authors.

Distinct Personas. The DAP model includes a regularization term specifically for encouraging personas with unique combinations of topics. We examined the top three topics associated with each persona. In the unregularized model, the 15 personas are only a mix of 6 different topics. In fact, a topic on "Weather" appears as a common topic for all 15 personas. On the other hand, the regularized DAP model's personas are a mix of 18 different topics. Further, the most frequently appearing topic is "Cancer (general)" (in 6 of 15 personas), which is appropriate given that approximately half of authors report cancer as a health condition.

Personas Reflect Coherent Health Journeys. In Figure 3 we show the top three topics evolving over time for selected personas. Labels for each topic are created manually based on words and journals most associated with the topic. Words most associated with each topic are listed in Tables 2. The persona plots in Figure 3 paint a compelling picture of common health journeys experienced by CB users.

Personas reflect broad trends, often encompassing a range of health journeys. Consider Persona 9, which reflects health journeys beginning with a physical element, such as physical therapy or a health issue taking a physical toll, followed by intensive care and attention to weight. Many Persona 9 authors begin physical therapy following an accident, or are caring for a premature baby or child with a congenital disorder. However, there are a number of rare disorders that follow Persona 9's pattern. For instance, one Persona 9 author writes about a family member with Guillain-Barré syndrome, a rare rapid-onset disorder in which the immune system attacks the nervous system resulting in muscle pain, weakness, and even paralysis. The syndrome often requires admittance to an intensive care unit, followed by rehabilitation – all common themes of Persona 9.

7 Conclusion

The Dynamic Author-Persona topic model is uniquely suited to modeling text data with a temporal structure and written by multiple authors. Unlike previous temporal topic models, DAP discovers latent personas — a novel component that identifies authors with similar topics trajectories. Our RVI algorithm further improves the DAP model's performance over competing models and results in the discovery of distinct personas. In evaluating the DAP model, we introduce the CaringBridge dataset: a massive collection of journals written by patients and caregivers, many of who face serious, life-threatening illnesses. From this dataset the DAP model extracts compelling descriptions of health journeys.

Acknowledgments

We thank reviewers for their valuable comments, University of Minnesota Supercomputing Institute (MSI) for technical support, and CaringBridge for their support and collaboration. The research was supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560.

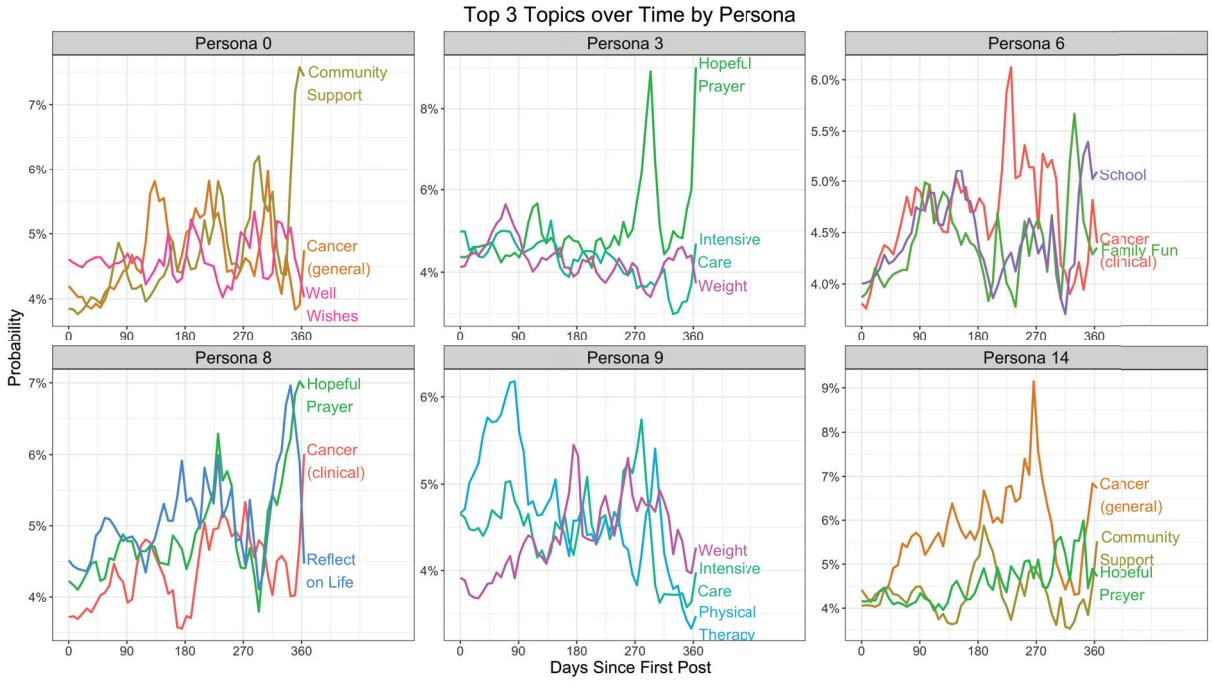

Figure 3: The unregularized DAP model finds compelling, unique personas corresponding to common health journeys experienced by CaringBridge users. The three most likely topics for personas are plotted over time. Results shown for six personas that highlight diversity in topic focus. Personas 0, 6, 8, and 14 highlight nuances in how an author writes about a topic like cancer. Personas 0 and 14 engage with their community, and are less clinical when writing about cancer. Persona 14's journals, however, are more religious and often include prayer. On the other hand, when discussing health, Personas 6 and 8 write about cancer using clinical terminology. When persona 6 is not sharing health updates the conversation is often on school, family, and celebrations. Whereas, persona 8's non-health updates are deep, reflective, and prayerful.

Community Support	Physical Therapy	Reflect on Life	Hopeful Prayer	Family Fun	Infection	Weather	School
family	therapy	life	god	christmas	blood	nice	school
friend	rehab	know	pray	play	infection	weather	shot
church	therapist	child	prayer	birthday	fluid	walk	go
thank	physical	never	lord	game	fever	lunch	appt
card	pt	love	bless	fun	antibiotic	cold	class
love	chair	year	please	kid	pressure	snow	tomorrow
service	speech	live	heal	party	kidney	outside	grandma
friends	progress	people	trust	year	iv	breakfast	teacher
support	move	cancer	peace	enjoy	lung	rain	home
gift	arm	moment	continue	dinner	clot	go	aunt

Cancer (clinical)	Cancer (general)	Intensive Care	Well Wishes	Hair Loss	Surgery	Bedtime	Weight
chemo	cancer	tube	dad	hair	surgery	sleep	weight
blood	treatment	breathe	mom	leg	surgeon	night	mommy
count	radiation	oxygen	everyone	wear	heart	bed	gain
bone	scan	lung	message	head	dr	wake	feed
marrow	chemo	feed	guestbook	look	office	nurse	daddy
platelet	tumor	x_ray	please	cut	op	say	bottle
round	oncologist	chest	prayer	knee	procedure	asleep	pound
clinic	dr	nurse	read	hat	cardiologist	_time_	feeding
transfusion	ct	vent	visit	wig	valve	room	oz
url	result	stomach	update	shave	ha	tell	milk

Table 2: Top 10 words associated with the most prevalent topics found by the DAP model ($\rho = 0.2$). Topic labels are selected manually in order to aid reference with Figure 3. The words _time_ and _URL_ refer to the result of text pre-processing steps for capturing common patterns like the time of day and website URLs, respectively.

References

Beaudoin, C. E., and Tao, C.-C. 2007. Benefiting from Social Capital in Online Support Groups: An Empirical Study of Cancer Patients. *CyberPsychology & Behavior* 10(4):587–590.

Blei, D. M., and Lafferty, J. D. 2006. Dynamic Topic Models. *International Conference on Machine Learning* 113–120.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2016. Variational Inference: A Review for Statisticians. *arXiv* arXiv:1601.00670v1 [stat.CO].

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4-5):993–1022.

Chang, J.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22* 288—-296.

Lafferty, J. D., and Blei, D. M. 2006. Correlated Topic Models. *Advances in Neural Information Processing Systems 18* 147–154.

McCallum, a.; Corrada-Emmanuel, a.; and Wang, X. 2005. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *NIPS'04 Workshop on'Structured Data and Representations in Probabilistic Models for Categorization* 1–16.

Mimno, D., and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 500–509.

Pathak, N.; DeLong, C.; Erickson, K.; and Banerjee, A. 2008. Social topic models for community extraction. *The 2nd SNA-KDD Workshop '08 (SNA-KDD'08)*.

Rodgers, S., and Chen, Q. 2005. Internet community group participation: Psychosocial benefits for women with breast cancer. *Journal of Computer-Mediated Communication* 10(4):1–27.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence* 487–494.

Steyvers, M.; Smyth, P.; Rosen-Zvi, M.; and Griffiths, T. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 306–315. ACM.

Wainwright, M. J., and Jordan, M. I. 2007. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning* 1(12):1–305.

Wang, X., and McCallum, A. 2006. Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* 424–433.

Wang, C.; Blei, D.; and Heckerman, D. 2008. Continuous Time Dynamic Topic Models. *Proc of UAI* 579–586.

Wei, X.; Sun, J.; and Wang, X. 2007. Dynamic Mixture Models for Multiple Time Series. *Ijcai* 2909–2914.

Wen, M., and Rose, C. P. 2012. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. *Proceedings of the 17th ACM international conference on Supporting group work - GROUP '12* 179.

Wen, K. Y.; McTavish, F.; Kreps, G.; Wise, M.; and Gustafson, D. 2011. From Diagnosis to Death: A Case Study of Coping With Breast Cancer as Seen Through Online Discussion Group Messages. *Journal of Computer-Mediated Communication* 16(2):331–361.