Mach Learn (2013) 92:431–455
DOI 10.1007/s10994-013-5348-5

Modeling individual email patterns over time with latent variable models

Nicholas Navaroli · Christopher DuBois · Padhraic Smyth

Received: 11 January 2013 / Accepted: 1 April 2013 / Published online: 1 May 2013

Abstract As digital communication devices play an increasingly prominent role in our daily lives, the ability to analyze and understand our communication patterns becomes more important. In this paper, we investigate a latent variable modeling approach for extracting information from individual email histories, focusing in particular on understanding how an individual communicates over time with recipients in their social network. The proposed model consists of latent groups of recipients, each of which is associated with a piecewise-constant Poisson rate over time. Inference of group memberships, temporal changepoints, and rate parameters is carried out via Markov Chain Monte Carlo (MCMC) methods. We illustrate the utility of the model by applying it to both simulated and real-world email data sets.

Keywords Email analysis · Community detection · Changepoint detection · Hidden Markov models · Poisson regression

1 Introduction

With the ubiquity of modern communication channels, such as text messaging, phone calls, email, and microblogging, there is increasing interest in the analysis of streams of user communication data over time. This paper focuses on analyzing egocentric network data over time (and more specifically, email histories) consisting of time series of counts of communication events between an individual and his or her social circle. As an example of this type

Editors: Zhi-Hua Zhou, Wee Sun Lee, Steven Hoi, Wray Buntine, and Hiroshi Motoda.

N. Navaroli (✉) · P. Smyth
Department of Computer Science, University of California, Irvine, Irvine, CA, USA
e-mail: nnavarol@ics.uci.edu

P. Smyth
e-mail: smyth@ics.uci.edu

C. DuBois
Department of Statistics, University of California, Irvine, Irvine, CA, USA
e-mail: duboisc@ics.uci.edu

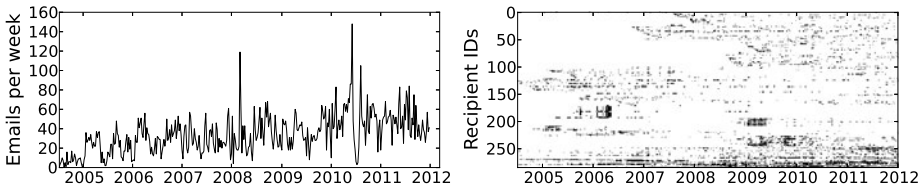


Fig. 1 Personal email communication data for one of the authors. *Left*: Total number of emails sent per week. *Right*: Points indicate an email sent to a particular individual (y-axis) at a particular time (x-axis)

of data, Stephen Wolfram recently described a personal data stream of emails, keystrokes, phone calls, and other activity that he has recorded over the past 20 years (Wolfram 2012), including a time series of one third of a million emails he has sent in that time. Our goal is to develop a statistical model that can summarize the major characteristics of such data: at what rates does an individual communicate? who does he or she communicate with? are there patterns among the recipients such as groups or clusters? and how do these patterns change over time?

As an example of the type of data of interest, Fig. 1 shows the weekly email communication patterns from several years of email history from one of the authors of this paper. The left plot shows the weekly rate of email over time. In the right plot the y-axis represents different recipients, with dots representing which recipients received emails during particular weeks. The patterns of communication are clearly non-stationary. As the sender transitioned over time through different universities, projects, collaborations, and social activities, the recipient patterns changed significantly over time, as did the overall communication rates. This data is not easy to summarize or interpret, and there are multiple different aspects of the data that are difficult to disentangle. For example, prior to 2008, recipients with IDs from 0 to 120 are rarely present. In addition, during the middle of 2010 there is a large peak followed by a sharp drop in communication, and it would be interesting to know which recipients are associated with this change in behavior. We will describe an unsupervised statistical learning approach that can explain such variations, in terms of both who we communicate with and the rate at which we communicate.

There are several potential applications of this type of model. For example, there is a broad consensus that there is significant room for improvement over current approaches to personal email management (Fisher et al. 2006; Whittaker et al. 2011; Wainer et al. 2011). This has motivated the development of a variety of visualization and clustering techniques as the basis for automated email management tools to help individual users to better understand and manage their email (Fisher 2005; Dredze et al. 2009a, 2009b; Koren et al. 2011; MacLean et al. 2011). A simple example is the Gmail "Got the wrong Bob?" feature (Roth et al. 2010) which learns about co-appearance patterns in email recipient lists and automatically suggests additional potential recipients to the sender. Further afield, in areas such as the social sciences, there is increasing interest in analyzing digital human communication data as a mechanism for investigating social theories about human behavior and interaction in the context of modern communication settings (e.g., Garton et al. 1997; Diesner et al. 2005; Alison Bryant et al. 2006; Butts 2008; Zenk et al. 2010; de Nooy 2011).

In this paper we focus on modeling recipient email counts over time, sent from a single account. The focus on lists of recipients (with respect to emails sent from the account of interest) is primarily a pragmatic one: it is a useful starting point and considerably simpler

than modeling both senders and receivers. Of the two, recipient lists are potentially of more interest in informing us about the individual since they are the result of active behavior by the individual, while sender counts are not directly so. It is natural to think of extensions to our approach here that can handle both sender and recipient information and such a model could be developed as an extension of the recipient-only model we present here.

Specifically, our proposed model consists of two interacting components:

1. Group structure among recipients is modeled by a mixed membership model, similar to that used in mixed membership models for social networks (Airoldi et al. 2008), and in topic modeling for text (Blei et al. 2003). This framework allows for modeling of recipients as members of multiple groups in a natural and parsimonious manner.
2. The daily number of emails sent over time is modeled via a set of independent piecewise-constant Poisson processes, one per group. The number of changepoints between Poisson segments for each group is handled by a non-parametric Dirichlet process, i.e., there are a potentially infinite number of changepoints and segments in the prior for the model, from which we infer a posterior distribution over a finite number of segments and changepoints given the observed data.

The primary contribution of this paper is a latent variable model that describes both group structure and non-stationary rate behavior for communication count data over time, with email data being the specific application focus. In Sect. 2 we discuss previous work on modeling count time series in email and other communication data, where the primary focus has been on segmentation and changepoint detection but without group structure. Sections 3 and 4 outline the model and our inference algorithms. In Sect. 5 we illustrate how the model works using simulated data. Section 6 illustrates the application of the model to real-world email data sets, demonstrating how the model can be used to better understand such data. Section 7 describes a set of experiments where we compare the proposed model to various baselines in terms of both the quality of the groups learned by the model and the predictive accuracy of the model. Section 8 contains discussion and conclusions.

2 Related work

Prior work that is relevant to our proposed approach can be broadly categorized into three areas: (1) models of email communication data, (2) segmentation of time series of count data, and (3) identification of group structure in dynamic social network data.

Earlier work on analysis of email time series has focused primarily on modeling of overall communication rates. For example, in a series of papers, Malmgren and colleagues have investigated a variety of bursty and non-homogeneous Poisson models to capture the overall rate at which an individuals send email (Malmgren et al. 2008, 2009). Earlier work in a similar vein applied Markov-modulated Poisson processes to telephone call and Web navigation data (Scott and Smyth 2003; Scott 2004). Our approach also uses latent piecewise-constant Poisson processes for modeling temporal variation in individual communication rates. We differ from prior work in that we show how the overall rate for an individual can be explained by a combination of (a) grouping patterns among recipients, and (b) time-varying rates for these groups—prior work focused on modeling just the overall rate for an individual, without recipient information.

In the broader context of segmentation of time series of count data, statistical learning approaches have been well studied. For example, Fearnhead (2006) models the number and location of the changepoints by placing priors over them and obtaining posterior samples. Chib (1998) models the time series with a finite-state left-right hidden Markov model

(HMM), such that changepoints are represented as latent state transitions. Our approach is similar to Chib (1998), but uses Dirichlet process priors in order to have a potentially infinite number of latent states, allowing for an arbitrary number of changepoints. A significant difference from previous work is that we do not detect changepoints in a single time series, but in the decomposition of the time series according to the (simultaneously learned) latent groups. Each latent group is associated with its own time series and changepoints. Our approach is also inspired by recent advances in using non-parametric Bayesian methods for other types of human communication, such as detecting speaker changes in audio recordings of meetings (Fox et al. 2011). These methods use a HMM with a flexible number of latent states. In this work, we similarly use non-parametric techniques for segmenting the time series for each of our K latent groups.

The third relevant strand of prior work is that of learning latent group structure from dynamic social network data. There is a large literature on this topic, including techniques based on optimizing a specific cost function or using statistical model-based approaches. One distinction among these methods is whether individuals are allowed to belong to one group or to several groups at any particular time. We take an approach that allows individuals to be members of multiple groups, akin to mixed membership models (Airoldi et al. 2008; Choi et al. 2012) and the work of McAuley and Leskovec (2012). In particular, we jointly model both the group memberships and the rate of events involving a particular group. This is similar in some respects to prior work on dynamic topic models, except that here we use a rate-based changepoint model to handle sudden changes in group rates over time, whereas the typical approach in dynamic topic models is to model changes in the probabilities of groups as a smooth function of time (e.g., Blei and Lafferty 2006; Wang et al. 2008). Also of relevance to our approach is prior work on community detection for dynamic social networks based on node clustering techniques, e.g., detecting clusters of nodes (communities) in a time-varying weighted graph. Such approaches include algorithms based on graph-coloring (Tantipathananandh et al. 2007) and clustering methods based on smoothed "snapshots" of a dynamic network (Xu et al. 2011). While one could in principle use these types of approaches for the grouping component of our model, we have chosen instead the mixed membership approach, which allows email recipients to belong to multiple groups at once. The probabilistic semantics of such a model allows us to learn and reason about both groups and communication rates in a coherent fashion.

3 The model

We begin in Sect. 3.1 by describing our approach to learning changepoints from time series data of counts using an infinite-state HMM, and then couple this with learning latent group structure in Sect. 3.2.

3.1 Modeling communication rates

Let N_t represent the total number of emails the user sends on day t. The set of variables $\{N_t : 1 \leq t \leq T\}$ define a stochastic process. We assume that $N_t \sim \text{Poisson}(\lambda_t)$, where λ_t is the rate at which the user sends emails on day t. Because λ_t is allowed to change across days, this type of process is usually referred to as a non-homogeneous Poisson process.

Our model assumes that the user communicates with K separate groups of people. Each email the user sends is sent to one of the K groups. We assume that the rate at which emails are sent to each group are independent Poisson processes, i.e., a change in the rate

at which emails are sent to one group does not affect the rate at which emails are sent to other groups. This assumption is clearly an approximation of what happens in practice—for example there may be exogenous (external) events, such as the user going on vacation, that affect most or all groups simultaneously. Nonetheless, we believe this independence model is a useful (and computationally efficient) place to start, allowing us to capture "first-order" group behavior—models allowing dependence between groups and/or shared dependence on exogenous events would be of interest as extensions of the simpler model we propose here.

Let $N_{k,t}$ represent the (unobserved) number of emails the user sends to group k on day t. We model $N_{k,t} \sim \text{Poisson}(\lambda_{k,t})$, where $\lambda_{k,t}$ is the rate at which the user sends emails to group k on day t. Because of our independence assumptions, N_t is the superposition of independent Poisson processes ($N_t = \sum_{k=1}^{K} N_{k,t} \sim \text{Poisson}(\lambda_t)$, where $\lambda_t = \sum_{k=1}^{K} \lambda_{k,t}$).

For the remainder of this subsection, we describe the model for time-varying communication rates for a single group, deferring discussion of how we learn the groups themselves to Sect. 3.2. We model a user's email rate to group k, $\{\lambda_{k,t} : 1 \leq t \leq T\}$, using a HMM. Under the HMM, the value of $\lambda_{k,t}$ is dependent on a latent state $s_{k,t}$, and the value of $s_{k,t}$ is dependent on $s_{k,t-1}$, the state of the previous day. Unique states represent different modes of activity between the user and recipient groups.

We define a *changepoint* to be a time t where the HMM transitions between different states ($s_{k,t} \neq s_{k,t+1}$). Changepoints will typically correspond to unobserved events throughout the user's history that change their communication rate with the group (such as vacations, research deadlines, changing schools, etc.). We define the single, contiguous interval of time between two adjacent changepoints to be a *segment*. Each segment represents a period of constant mean activity for the user with respect to a particular group.

Traditional HMMs have a finite number of states, limiting the modes of activity a user can have. Here we allow the HMM to have a countably infinite number of states, where only a finite subset of those states are ever seen given the observed data (similar to Beal et al. 2002). We enforce the restriction that the HMM cannot transition to previously seen states (known as a *left-to-right* HMM), ensuring that each unique state spans a single interval of time.[1] We model such a HMM by placing separate symmetric Dirichlet priors over each row of the transition matrix. As the number of latent states tends to infinity, these priors converge in distribution to Dirichlet processes (Neal 2000). A property of Dirichlet processes is that, after integrating out the parameters for the HMM transition matrix, the transition probabilities between states become:

$$P(s_{k,t}|s_{k,-t},\gamma,\kappa) = \begin{cases} \frac{V_t+\gamma}{V_t+\gamma+\kappa} & \text{if } s_{k,t} = s_{k,t-1}, \\ \frac{\kappa}{V_t+\gamma+\kappa} & \text{if } s_{k,t} \text{ is a new state}, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where γ and κ are adjustable parameters, $s_{k,-t} = \{s_{k,t'} : t' \neq t\}$ is the set of all other states (not just the previous state, since the integration of the transition matrix introduces dependencies between all latent states), and $V_t = \sum_{t'=2}^{t-1} \delta(s_{k,t'} = s_{k,t-1})\delta(s_{k,t'-1} = s_{k,t-1})$ is how

[1] The alternative approach of allowing the state transition matrix to be unconstrained (i.e., allowing the HMM to return to earlier states, as in Fox et al. 2011) is also certainly feasible, and has the advantage that segments could share parameters by representing recurring states and rates. We did not pursue this approach primarily for computational reasons since inference in such a model is significantly more complex than in the proposed changepoint left-to-right model.

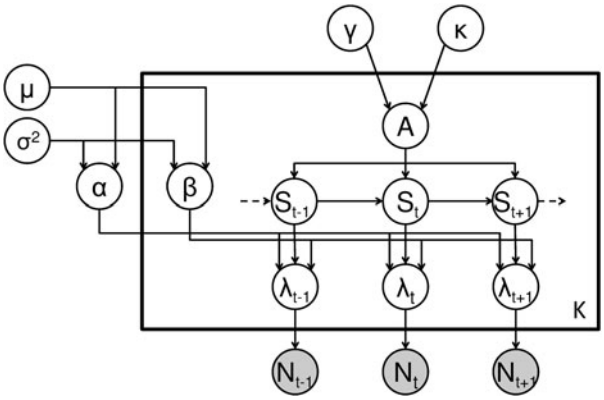


Fig. 2 Graphical representation of the K HMMs associated with the K groups. A represents HMM transition matrices

long the HMM has been in state $s_{k,t-1}$ up to time t. Appendix A contains additional discussion of the sensitivity to segment lengths to γ and κ.

The other dependence to model in the HMM is how group k's rate at time t depends on its latent state $s_{k,t}$, namely $\lambda_{k,t}|s_{k,t}$. We use Poisson regression to model the log of these rates, i.e., $\log \lambda_{k,t} = X_{k,t}^T \theta$, where $X_{k,t}^T$ is a set of features for day t and θ is a vector of regression parameters. We construct $X_{k,t}^T$ and θ such that $\log \lambda_{k,t} = \beta_{k,s_{k,t}}$, where $\beta_{k,m}$ is the log of the rate that the user is sending emails to group k while in time segment m (corresponding to state m of the HMM). $X_{k,t}$ is a binary vector indicating the latent state of the HMM on day t, and $\theta = [\beta_{k,1}, \beta_{k,2}, \ldots, \beta_{k,M_k}]^T$, where M_k is the number of unique states. Because we are modeling $\lambda_{k,t}$ with Poisson regression, we can also include other features (which may or may not depend on group k). For example, in the results in this paper we include *day-of-week* effects:

$$\log \lambda_{k,t} = \beta_{k,s_{k,t}} + \alpha_{d(t)}, \tag{2}$$

where $d(t) \in W$ represents different days of the week. We can, for example, use $W = \{0, 1\}$ to represent weekdays and weekends. Having this configuration allows the overall number of emails sent by the user to vary between weekdays and weekends, while having the relative emailing rates for each group remain unchanged. As an example, consider a user that only sends emails on weekdays. The α_{weekend} regression term would have a large negative value, forcing $\lambda_{k,t} \approx 0$ for every group on weekends. In the results in this paper we use $W = \{0, \ldots, 6\}$, allowing overall activity to be modulated for each day of the week individually.

Figure 2 shows the overall structure of the HMM in the form of a graphical model, with a different $\lambda_{k,t}$ for each group k. Since the λ's are deterministic functions of α and β (using (2)) they could be omitted from the depiction of the graphical model, but are included for clarity.

3.2 Modeling recipient groups

We discuss next how to model the K different groups that the user interacts with, where a group is defined as a distribution over the possible sets of recipients for an email. Intuitively the different groups should account for different relationships that are reflected in the individual's email time series, such as familial and friend relationships, organizational relationships, recreational activities, and so forth.

We assume each email is sent to one of the K latent groups. Let $z_{t,n}$ represent the latent group that email n on day t was sent to, and $y_{t,n} \in [0, 1]^R$ a binary vector indicating which of

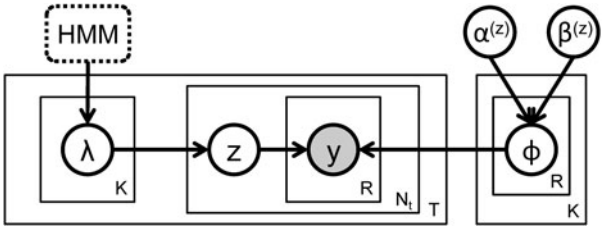


Fig. 3 Graphical representation of the K groups. The "HMM" supernode is a compact representation of Fig. 2

the R possible recipients are in the email. Given the latent group, the recipients of the email are selected independently with probability $\phi_{k,r}$, i.e., a conditionally-independent Bernoulli model. The generative model is

$$\phi_{k,r}|\alpha^{(z)},\beta^{(z)} \sim \text{Beta}\big(\alpha^{(z)},\beta^{(z)}\big), \qquad y_{t,n}|\phi, z_{t,n} \sim \prod_{r=1}^{R} \text{Bernoulli}(\phi_{z_{t,n},r}), \tag{3}$$

where $\alpha^{(z)}$ and $\beta^{(z)}$ are the parameters of an independent set of Beta priors over the individual recipient probabilities. Under this model the expected number of recipients for an email sent to group k is $\sum_r \phi_{k,r}$.

This modeling of the latent group indicator variables $z_{t,n}$ is a key aspect of the model; the distribution of $z_{t,n}$ introduces dependencies between the K separate HMMs from Sect. 3.1 and the generative model of email recipients in (3). The discrete probabilities over the latent variables are a function of the daily rates $\lambda_{k,t}$, the rate at which the user is sending emails to group k on day t. Because $\{\lambda_{k,t} : 1 \le t \le T\}$ is a Poisson process and the time between emails for a particular group follows an exponential distribution, it is straightforward to show (using standard properties of exponential random variables; e.g., Ross 2007) that the probability of an email on day t is sent to group k can be written as

$$P\big(z_{t,n} = k \mid \{\lambda_{k',t} : 1 \le k' \le K\}\big) = \frac{\lambda_{k,t}}{\sum_{k'=1}^{K} \lambda_{k',t}}. \tag{4}$$

Figure 3 shows the graphical model for representing the group aspect of the model, as described in this section. In the interest of interpretability, all the HMM variables described in Fig. 2 (except for λ and N) are combined into a single supernode.

4 Parameter inference

We use MCMC techniques to learn the parameters of our model from observed data. In particular, we use Gibbs sampling to iteratively sample each of the variables from their full conditional distributions. These conditional distributions can be derived using the graphical models in Figs. 2 and 3, since the joint distribution over all parameters (which the conditional probabilities are proportional to) factors according to the graphical model. We outline the sampling equations for each variable in the following subsections—more complete derivations are provided in Appendix B. To keep the notation simple, variables without subscripts denote the set of all variables that can be indexed by it, e.g., $\lambda = \{\lambda_{k,t} : 1 \le k \le K, 1 \le t \le T\}$. We also use Θ to denote the set of all model parameters and variables.

4.1 Sampling the latent groups

By taking advantage of the conjugacy between the Beta and Bernoulli distributions, we can integrate out the membership probabilities ϕ analytically. The conditional distribution for sampling $z_{t,n}$ given all other variables is

$$P(z_{t,n} = k|\Theta \setminus z_{t,n}) \propto P(z|\lambda) \int P(y|z,\phi) P\big(\phi|\alpha^{(z)}, \beta^{(z)}\big) d\phi$$
$$\propto \lambda_{k,t} \prod_{r=1}^{R} \left(\frac{(c_{1,k,r}^{-(t,n)} + \alpha^{(z)})^{y_{t,n,r}} (c_{0,k,r}^{-(t,n)} + \beta^{(z)})^{(1-y_{t,n,r})}}{c_{1,k,r}^{-(t,n)} + \alpha^{(z)} + c_{0,k,r}^{-(t,n)} + \beta^{(z)}} \right),$$

where

$$P(z|\lambda) = \prod_{t=1}^{T} \prod_{n=1}^{N_t} P\big(z_{t,n}|\{\lambda_{k,t} : 1 \le k \le K\}\big),$$
$$P(y|z,\phi) = \prod_{t=1}^{T} \prod_{n=1}^{N_t} P(y_{t,n}|\phi_{z_{t,n}}),$$
$$P\big(\phi|\alpha^{(z)}, \beta^{(z)}\big) = \prod_{k=1}^{K} \prod_{r=1}^{R} P\big(\phi_{k,r}|\alpha^{(z)}, \beta^{(z)}\big),$$
$$c_{0,k,r}^{-(t,n)} = \sum_{t' \neq t} \sum_{n' \neq n} \delta(z_{t',n'} = k)(1 - y_{t',n',r}),$$
$$c_{1,k,r}^{-(t,n)} = \sum_{t' \neq t} \sum_{n' \neq n} \delta(z_{t',n'} = k) y_{t',n',r}.$$

$c_{1,k,r}^{-(t,n)}$ is the number of emails sent to group k that recipient r appears in, ignoring email n on day t. $c_{0,k,r}^{-(t,n)}$ is the number of emails sent to group k that recipient r does not appear in, ignoring email n on day t. The sampler samples a new value for each $z_{t,n}$ in sequential order, starting with the first email sent. Once a new value for $z_{t,n}$ is sampled, the statistics for how often recipients appear (and do not appear) in emails sent to each group are updated. The sampling process for these variables, along with their conditional distribution derivations, is similar to that of standard collapsed Gibbs sampling algorithms for latent variable models for discrete data, for example as used for LDA.

Throughout the paper we place a Beta prior, with $\alpha^{(z)} = 0.01, \beta^{(z)} = 0.01$, over each recipient probability $\phi_{k,r}$. This prior reflects the belief that recipients that are active in a group should appear in most emails sent to the group, and recipients that are not active in a group should rarely appear.

4.2 Sampling the regression parameters

We place a Normal(μ, σ^2) prior on the regression parameters β and α, and use $\mu = 0, \sigma^2 = 1$ for the results in this paper. Unlike Sect. 4.1, this prior is non-conjugate, making it intractable to sample directly from the conditional distribution. We can produce a valid Gibbs sample by instead sampling from the unnormalized log conditional distribution, via a technique

known as slice sampling (Neal 2003). These unnormalized log conditional distributions for sampling the regression parameters are

$$\begin{aligned}\log P(\alpha_w|\Theta\setminus\alpha_w) &\propto \log\bigl(P\bigl(\alpha_w|\mu,\sigma^2\bigr)P(N|\alpha,\beta,s)\bigr)\\ &\propto -\frac{(\alpha_w-\mu)^2}{2\sigma^2}+\alpha_w\sum_{t:d(t)=w}N_t-e^{\alpha_w}\left(\sum_{t:d(t)=w}\sum_{k=1}^{K}e^{\beta_{k,s_{k,t}}}\right),\\ \log P(\beta_{k,m}|\Theta\setminus\beta_{k,m}) &\propto \log\bigl(P\bigl(\beta_{k,m}|\mu,\sigma^2\bigr)P(N|\alpha,\beta,s)P(z|\alpha,\beta,s)\bigr)\\ &\propto -\frac{(\beta_{k,m}-\mu)^2}{2\sigma^2}-\sum_{t:s_{k,t}=m}e^{\alpha_{d(t)}}e^{\beta_{k,m}}+g_{k,m}\beta_{k,m},\end{aligned}$$

where

$$\begin{aligned}P(N|\alpha,\beta,s)&=\prod_{t=1}^{T}P(N_t|\alpha,\beta,s),\\ P(z|\alpha,\beta,s)&=P(z|\lambda)\quad\text{(in Sect. 4.1)},\\ g_{k,m}&=\sum_{t=1}^{T}\sum_{n=1}^{N_t}\delta(z_{t,n}=k)\delta(s_{k,t}=m).\end{aligned}$$

$g_{k,m}$ is the number of times an email was sent to group k when that group was in segment m. Note that the emailing rates, λ, are deterministic functions of α and β (using (2)).

4.3 Sampling the HMM hyperparameters

Instead of fixing the Dirichlet process hyperparameters γ and κ, we place priors on them and learn them from the data. We define priors over their ratio $r=\frac{\gamma}{\gamma+\kappa}$ and magnitude $m=\gamma+\kappa$. The ratio r represents the probability of staying in a newly visited state, and the magnitude m represents the strength of the prior. As priors we use $m\sim\text{Gamma}(k_g,\theta_g)$ and $r\sim\text{Beta}(\alpha^{(r)},\beta^{(r)})$. As with the regression parameters, these priors are non-conjugate, so we use slice sampling over the unnormalized log conditional probability. We first sample m, which deterministically updates γ and κ. We then sample r, which updates γ and κ a second time. The conditional probabilities depend only on the priors and the HMM latent state probabilities:

$$\begin{aligned}P\bigl(m|\Theta\setminus\{\gamma,\kappa\}\bigr)&\propto P(m|k_g,\theta_g)P(s|\gamma,\kappa),\\ P\bigl(r|\Theta\setminus\{\gamma,\kappa\}\bigr)&\propto P\bigl(r|\alpha^{(r)},\beta^{(r)}\bigr)P(s|\gamma,\kappa),\end{aligned}$$

where

$$P(s|\gamma,\kappa)=\prod_{k=1}^{K}\prod_{t=1}^{T}P\bigl(s_{k,t}\mid\bigl\{s_{k,t'}:t'\neq t\bigr\},\gamma,\kappa\bigr).$$

We use a Gamma prior, with $k_g=0.5$, $\theta_g=20$, over the magnitude m and a Beta prior, with $\alpha^{(r)}=100$, $\beta^{(r)}=1$ over the ratio r. The probability $P(s_{k,t}|\{s_{k,t'}:t'\neq t\},\gamma,\kappa)$ is calculated using (1).

4.4 Sampling the segments

For each day t and group k we sample the latent state $s_{k,t}$ conditioned on (a) all other latent states for group k, (b) the latent states for other groups on day t, and (c) the emails sent on day t. We only sample $s_{k,t}$ where $s_{k,t-1} \neq s_{k,t+1}$, due to the restriction that the HMM cannot transition back to previous states. If $s_{k,t}$ is sampled, its possible values are the previous state $s_{k,t-1}$, the next state $s_{k,t+1}$, or a brand new state. The prior probability of entering a new state is proportional to the HMM hyperparameter κ. The conditional probability for sampling $s_{k,t}$ is

$$P(s_{k,t}|\Theta \setminus s_{k,t}) \propto P(N|\lambda)P(z|\lambda)P(s|\gamma,\kappa)$$

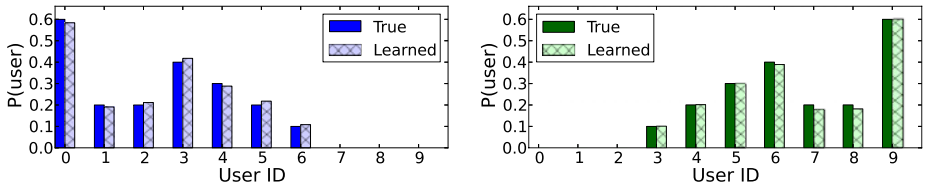
Note that in order to calculate the probability of $s_{k,t}$ being a brand new state, we first need a new β regression parameter for that new state. We sample the value of $s_{k,t}$ by first sampling this new regression parameter from its prior distribution, then using this new parameter in the above equation. This is an example of sampling using auxiliary variables (Neal 2000), where to sample from $p(x)$, we sample from a distribution $p(x,\xi)$ whose marginal distribution is $p(x)$. The auxiliary variable ξ is then discarded. In our case, $x = \Theta$, the set of all model parameters, and ξ represents the newly sampled β parameter. If $s_{k,t}$ is a singleton state (it is a segment of length one), it is possible for the segment to become "absorbed" into one of its neighboring segments during sampling. When this occurs, the corresponding β regression parameter no longer represents a segment. As is common in the application of Dirichlet processes, such parameters are discarded.

5 An illustrative example using synthetic data

As an illustration of the fitting procedure we created a synthetic data set with $K = 2$ groups, $T = 350$ days, and $R = 10$ possible recipients. The dark bars in Fig. 4 show the membership probabilities of the two groups; each group has three recipients unique to that group, with the remaining four recipients common across both groups. The communication rates for each group (the β regression parameters) are fixed, and are shown in Fig. 5(a). The rates are defined so that each group is dominant during different periods of time. The interaction with group 0 (represented with solid blue lines) changes on days 100 and 300, where the emailing rate changes. Similarly, the interaction with group 1 (represented with dashed green lines) changes on days 50, 120, and 210. The values for the α regression parameters are seen in Fig. 6(a), and are set so that the user's activity on weekends reduces by 40 %. The emailing rates to both groups, taking weekly patterns into effect, is seen in Fig. 5(b).

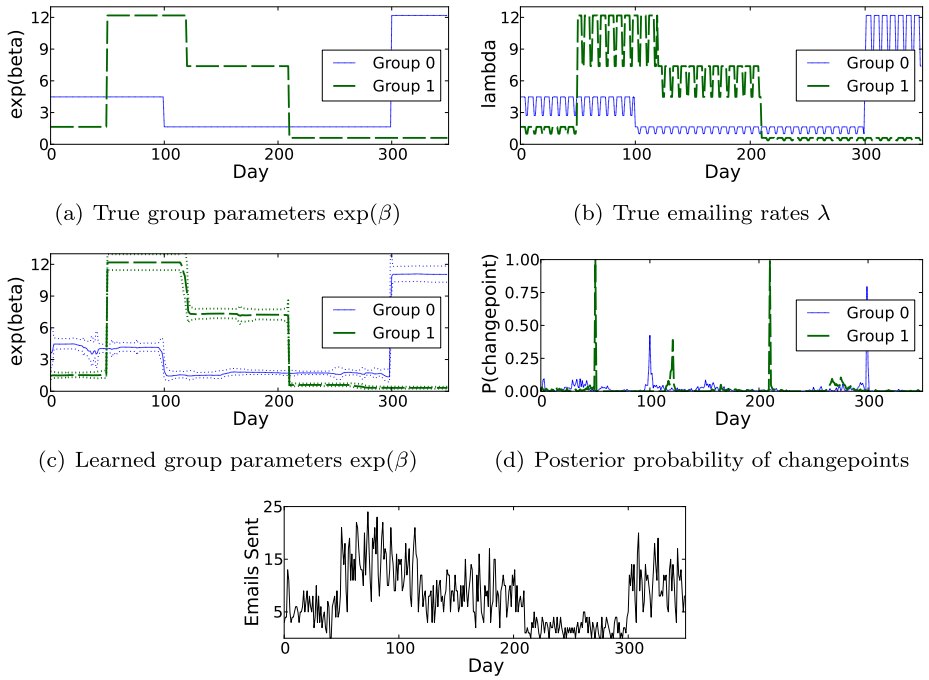
Given the parameters of the model, emails are simulated by first simulating the total number of emails sent for each day; $N_t \sim \text{Poisson}(\sum_{k=1}^{K} \lambda_{k,t})$. Figure 5(e) shows the sampled values for N_t. On day t, for each of the N_t emails, we simulate which group the email was sent to; $z_{t,n} \sim \text{Categorical}(\{p_k : 1 \leq k \leq K\})$, where $p_k \propto \lambda_{k,t}$. Once the group for an email is determined, recipients of the email are selected by iterating through each possible recipient r, and including them in the recipient list with probability $\phi_{z_{t,n},r}$.

To learn the parameters of the model, we iteratively sample the parameters of our model as described in Sect. 4. After collecting 2100 samples, the first 100 are discarded as burn-in, and every tenth sample after that is kept (for a total of 200 samples). The latent states for each HMM are initialized such that every $s_{k,t}$ is its own unique state. In other words, each group has 350 segments, each of length one day. The regression and Dirichlet process parameters



(a) Group 0 (b) Group 1

Fig. 4 True (*solid bars*) and learned (*cross-hatched bars*) membership probabilities for synthetic data



(a) True group parameters exp(β) (b) True emailing rates λ

(c) Learned group parameters exp(β) (d) Posterior probability of changepoints

(e) Number of emails sent each day

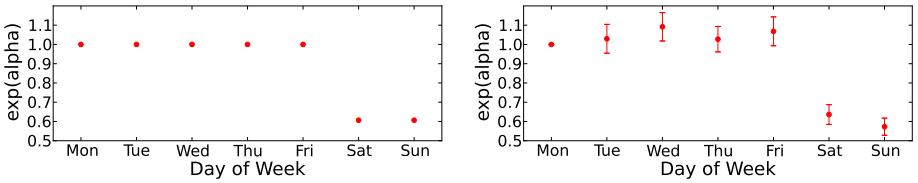
Fig. 5 True and learned parameters for the synthetic email user

are initialized to a sample from their prior distributions. The latent group variables $z_{t,n}$ are initialized uniformly at random between the two groups.

The cross-hatched bars in Fig. 4 show the learned groups after sampling. Membership probabilities are integrated out during sampling—using the set of latent z variables from the sample that produced the largest log-likelihood (maximum a posteriori (MAP) estimate), we estimate these probabilities as follows:

$$\hat{\phi}_{k,r} = \frac{\alpha^{(z)} + \sum_{t=1}^{T} \sum_{n=1}^{N_t} \delta(z_{t,n} = k)\delta(y_{t,n,r} = 1)}{\alpha^{(z)} + \beta^{(z)} + \sum_{t=1}^{T} \sum_{n=1}^{N_t} \delta(z_{t,n} = k)}.$$

In words, $\phi_{k,r}$, the probability of recipient r appearing in an email sent to group k, is estimated as the fraction of emails including recipient r that were sent to group k. Figure 5(c)



(a) True weekly parameters $\exp(\alpha)$ (b) Learned weekly parameters $\exp(\alpha)$

Fig. 6 True and learned day-of-week regression parameters for the synthetic email user

shows the average β regression parameters for both groups across the 200 samples, with dashed lines showing one standard deviation. Comparing this plot to Fig. 5(a), the model is able to learn the correct values of the regression parameters, even when the email rates are relatively small for both groups. The model was also able to learn the correct α parameters, shown in Fig. 6(b). A separate α parameter was estimated for each day of the week, with the points representing the mean across the samples, and the error bars showing one standard deviation. Note that the α parameter for the first day of the week has zero variance. This parameter is fixed, allowing the other regression parameters to scale accordingly. Figure 5(d) shows the posterior probability of changepoints for the two groups. The posterior probability of a changepoint on day t for group k is estimated as the fraction of samples where $s_{k,t} \neq s_{k,t+1}$. For both groups, the posterior probability peaks at the true changepoints in the simulated data. This indicates that the model is able to both decompose the overall email activity appropriately across the two groups, and learn when the behavior with respect to each group changes.

The results in this section are intended to be illustrative and demonstrate that the learning algorithm for the model is behaving as expected—the primary interest of course is what happens when we use the model on real data, which we discuss in the next section.

6 Exploratory analysis of email data

In this section we analyze data from the email accounts of the authors of this paper. For each author's email account, a script downloaded and parsed all the emails sent by that author. Recipients that received less than 10 emails were removed. Each email was represented by a time-stamp and a list of 1 or more integers representing anonymized IDs of recipients of that email.

For each data set, we learn the parameters of the model using $K = 2\sqrt{R}$ groups (rounded to the nearest integer), where R is the total number of unique recipients the user emailed over the time period. We found this to be a simple and effective heuristic for automatically choosing the number of groups to learn. An extension to the model would be to learn K automatically from the data, e.g., using Bayesian non-parametric techniques.[2] We use the same initialization of model parameters and configurations of the hyperparameters as the

[2]We chose not to follow this approach because of the additional complexity required to initialize a HMM for a new group, to average across samples to estimate emailing rates, and so on. Additionally, standard CRP modeling techniques are not directly applicable here as the distribution of latent group variables depend on the emailing rates for each group.

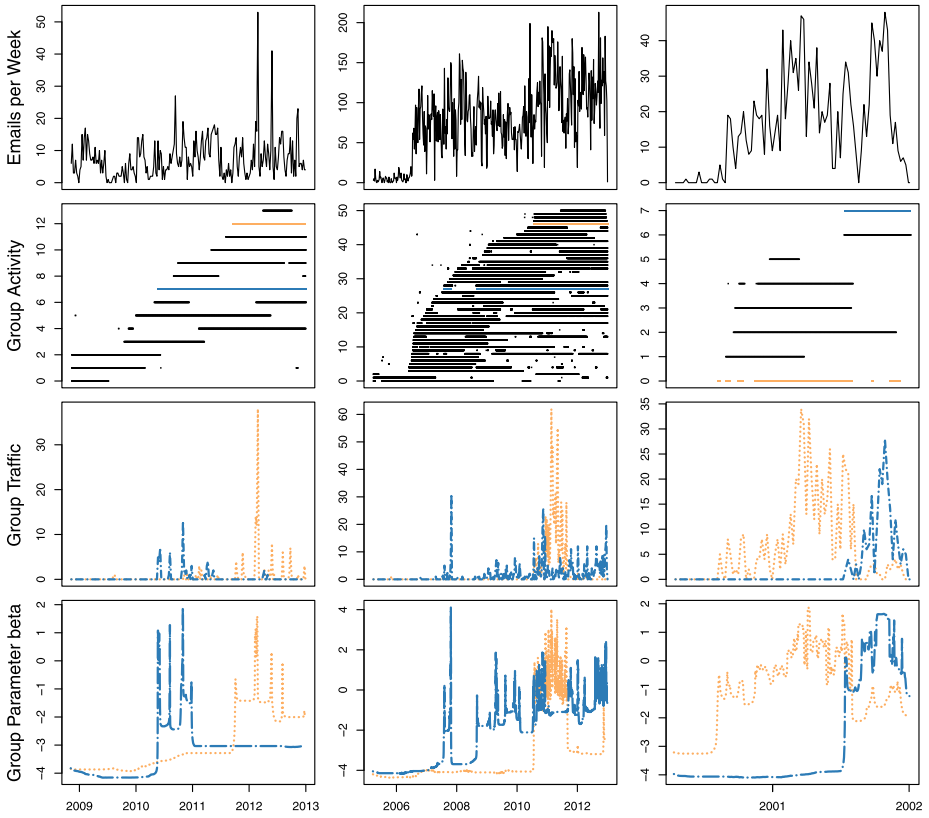


Fig. 7 Exploratory analysis from fitting the model to real-world email data. Each column corresponds to the email history of a specific individual. *First row*: Observed weekly email volume. *Second row*: greater than average β values for each learned group. *Third row*: Number of emails per week assigned to two chosen groups, highlighted throughout as *dashed* (*blue*) and *dotted* (*orange*). *Fourth row*: The learned parameters β for the chosen groups. See text for details

synthetic data in the previous section. A total of 21000 Gibbs samples are produced, discarding the first 1000 as burn-in, and keeping every 20th sample after that (for a total of 1000 samples). Model parameters are estimated from these samples as described in Sect. 5.

Figure 7 shows the learned parameters of the model for different email users. The top row of plots consist of the number of emails sent each week by each user. The second row shows the time intervals for which there was significant activity between the user and each group, where activity is represented as horizontal bars along the x-axis. These intervals were determined by thresholding the value of the β regression parameters for each group. The third row of plots show the number of emails the user had sent to two particular groups over time. The last row shows the learned β regression parameters for these two groups, again across time. The three columns correspond to three different users. The first and second columns correspond to the email accounts of two of the authors of this paper and the third column corresponds to one of the most active users from the Enron corpus (Klimt and Yang 2004).

The left column of Fig. 7 shows the learned parameters of the model for one of the authors of this paper. There is evidence in the group activity that a significant change oc-

curred in this user's behavior, around the middle of 2009. This changepoint corresponds to the user transitioning between two institutions: old connections faded and new connections were formed when moving from one location to another. The bottom two figures in the first column show email activity between this user and two learned groups, corresponding to two separate research projects the user participated in. The two research projects have a small number of common participants, illustrating that the model can use co-appearance information to successfully disambiguate groups that share common subsets of recipients. Spikes in behavior between the user and two groups correspond to different deadlines, e.g., progress reports, paper submissions, presentations, etc.

The center column of Fig. 7 shows what the model learns for a different author of this paper—this author sends considerably more emails than the user in the first column. Email activity is low for the first year as the user was experimenting with new email client software, followed by a sudden change and increase in activity as the user switched all email activity to the new client. The second row shows a gradual accumulation of groups over time (more horizontal bars start to appear), with groups that the author communicates with on a regular basis as well as groups that are only active for specific intervals of time.

The bottom two plots of this column show the traffic and estimated β parameters for two of the learned groups. The dashed (blue) group corresponds to a project where the author is a principal investigator for a large proposal and the recipients are six other faculty members at a number of different institutions. There is increasing activity in mid-2007 due to proposal preparation, then a spike at the proposal deadline itself in late 2007, followed by a quiet period until early 2008 when the project was funded. The group activity is then characterized by relatively low levels of activity for the next few years as the project proceeded, punctuated by spikes of more intense activity once or twice a year around the times of project review meetings. The dotted (orange) line shows a different group of about 15 individuals involved in the organization of a large conference. The activity of this group ramped up in mid 2010 as the author became involved in conference organization, followed by roughly 12 months of relatively high activity until the actual conference in summer 2011.

The third column in Fig. 7 illustrates results for an active user from the Enron corpus. This user's email activity increased in early 2001 and there appears to have been a significant change in recipient groups near the middle of the year—activity for some groups stop while new activity begins for others.

7 Experiments

In this section we describe two sets of experiments. In Sect. 7.1 we explore the quality of the learned groups produced by our model, and in Sect. 7.2 we compare the predictive performance of our model with different baselines. The email accounts used in these experiments include those from the authors of the paper, in addition to the five Enron users who sent the largest number of emails. As in Sect. 6, the model and the baselines learn $K = 2\sqrt{R}$ groups for each email user. The parameters are learned by collecting a total of 1100 samples, with the first 100 samples discarded for burn-in and every tenth sample kept after that, leaving a total of 100 samples.

7.1 Quality of learned groups

To measure the quality of the learned groups, we define a *coherence* metric that measures how often recipients co-appear in the emails assigned to a particular group, k:

$$s_k = \frac{1}{C}\left(\sum_{i=1, j\geq i}^{R_k} P(r_i)P(r_j)P(r_i, r_j)\right),$$

where $C = \sum_{i=1, j\geq i}^{R_k} P(r_i)P(r_j)$ is a normalization constant, R_k is the number of unique recipients across emails assigned to this group, $P(r_i)$ is the probability of recipient i appearing in an email sent to the group, and $P(r_i, r_j)$ is the probability of recipients i and j co-appearing in an email sent to the group. The $P(r_i)$, $P(r_j)$, and $P(r_i, r_j)$ terms are estimated empirically by counting how often recipient i, recipient j, and both i and j, appear in recipient lists for emails assigned to group k, where the assignment of emails to group k is based on the latent z variables from the highest log-likelihood sample obtained from the Gibbs sampler.

The intuition behind this coherence measure s_k is it rewards putting pairs of individuals into the same group who often co-appear in the same emails—conversely, it penalizes putting pairs of recipients into the same group who rarely co-appear in a recipient list. Coherence is maximized ($s_k = 1$) when all pairs of individuals (in the set of R_k) co-appear in all emails assigned to that group. As the co-appearance graph of recipients becomes less well-connected, s_k decreases. Consider for example a group k with $R_k = 4$ individuals labeled A, B, C, and D. When every email contains all group members, the group is assigned the maximum coherence value of 1. A coherence score of less than 1 will be assigned if emails are only sent to pairs of recipients, e.g., involving $\{A, B\}$, $\{B, C\}$, $\{C, D\}$, etc. Even smaller scores are assigned to groups when some subsets of individuals never co-appear in an email, e.g., emails involving $\{A, B\}$ and $\{C, D\}$ but never $\{A, C\}$. We can compute an average coherence score s for all groups by taking a weighted sum of the s_k's with weights corresponding to the fraction of the total number of emails assigned to that group.

For this experiment we trained our model 20 times, each time randomly initializing the model parameters for the sampler. For measuring the coherence of the learned groups we calculate scores based on the latent group variables z from the sample that produced the largest log-likelihood, for each set of learned parameters. The mean and standard deviation of the overall coherence score s is reported across the 20 sets of learned parameters.

We compare the coherence of our learned groups with those learned in four baseline models. The first baseline is identical to our proposed model, except that we use a multinomial model for the membership probabilities with $\sum_r \phi_{k,r} = 1$ (as in Navaroli et al. 2012). A multinomial likelihood is placed over email recipients for this baseline, as opposed to a product of independent Bernoulli trials. The single segment baseline is restricted to only have a single time segment for each group and, thus, no time variation in the relative rates of groups. This baseline can be viewed as a probabilistic clustering of the recipients and the (constant) rate of emails sent to each group determine the cluster mixing coefficients. We implemented both multinomial and independent Bernoulli definitions versions of this baseline. The final baseline is a Uniform model, where the group assignments for emails $z_{t,n}$ are chosen uniformly at random among the K groups.

Figure 8 shows the coherence scores for each author and the Enron users. The y-axis is the coherence s for each model, averaged across the 20 separately-learned parameters, with error bars corresponding to one standard deviation. For each user, the coherence of

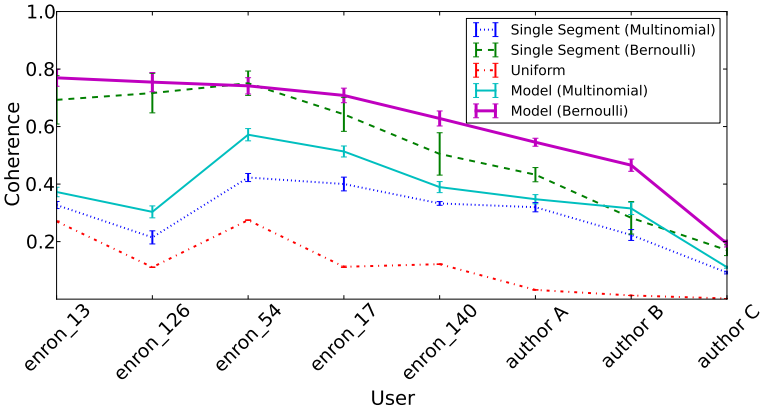


Fig. 8 Coherence scores (y-axis), averaged across 20 sampling initializations, for each email user (x-axis)

the groups learned by our model is higher than those learned by the four baselines. It is interesting to compare the results between the multinomial and independent Bernoulli definitions for groups in our model. The multinomial likelihood for recipients in an email is the product of membership probabilities for those recipients. Because this likelihood only places probability on recipients that appear in the email, it is not lowered by the absence of a high-probability recipient in an email. This allows for groups to be created that consist of multiple disjoint sub-groups. In contrast, the independent Bernoulli model produces a lower likelihood if a high-probability recipient is not in an email, making it harder for emails that are sent to disjoint sets of recipients to be merged into the same group.

7.2 Predictive performance

In this section we measure the predictive performance of our model when emails are removed uniformly at random from the training data. The parameters of the model are learned from the training data, ignoring the removed emails.[3] The predictive performance of the model and baselines on missing data is evaluated using the test log-likelihood:

$$LL_{\text{test}} = \sum_{t=1}^{T} \sum_{n:\text{missing}} \log\left(\sum_{k=1}^{K} P(z_{t,n} = k|\lambda, \phi) P(y_{t,n}|\phi_k) \right),$$

where the second sum is over the emails on day t that are in the test data set.

In our experiments below we generated 20 training and test data sets in this manner, randomly putting 20 % of emails in the test set each time, and computing the log-likelihood for each test set. For each training data set, the recipient probabilities ϕ_k and group emailing rates $\lambda_{k,t}$ are estimated using the method described in Sect. 5, using the samples collected from the training data set.

The predictive power of our model is compared against four baseline approaches. The *uniform* baseline consists of a single group, whose membership probabilities ϕ_r are equal

[3] One could also explicitly model the missing data by averaging over the missing information during MCMC sampling—however this would require a much more complex sampling algorithm so we opted for the simpler approach of ignoring missing data during training.

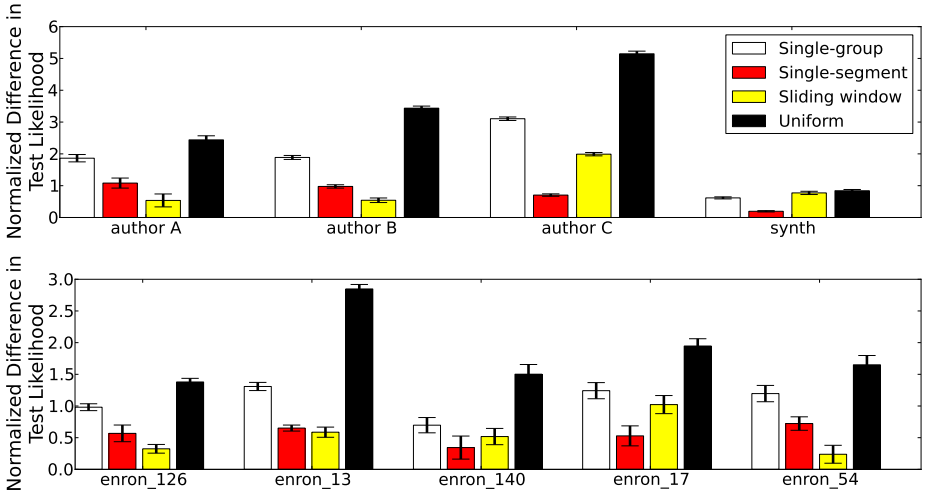


Fig. 9 Predictive results from a missing data task for different users

across all possible recipients and is set such that the expected number of recipients in an email matches that of the training data set. The *single-group* baseline corresponds to a maximum likelihood of an independent Bernoulli model over possible recipients, where ϕ_r is equal to the fraction of emails that recipient r appears in. The *sliding window* baseline is similar to the single-group baseline, except that the recipient probabilities are based on local time-windows, allowing the probabilities to adapt to changes over time in recipient likelihood. We evaluated different sized windows up to two months and used the one that gave the best results in the data reported here. The *single segment* baseline is identical to that in Sect. 7.1. For the first three baselines, only one group exists on any given day; the test log-likelihood of these baselines reduce to

$$LL_{\text{test}} = \sum_{t=1}^{T} \sum_{n:\text{missing}} \log P(y_{t,n}|\phi).$$

Figure 9 shows the results for the four baselines, relative to our model, across the authors and Enron users. The y-axis is the average difference in test log-likelihood between our independent Bernoulli model and each baseline, averaged over the 20 test data sets, where larger positive differences mean that the model outperformed the baseline. Each test log-likelihood score was normalized by dividing by the total number of missing emails for which predictions were made for that user. Authors A, B, and C correspond to several years of email data from each of the three authors of this paper, and the "synth" user is the synthetic data set described in Sect. 5. This plot shows results obtained with 20 % of emails missing at random—almost identical results were obtained with other fractions of missing data (not shown).

The results in Fig. 9 show that our proposed model is systematically more accurate in predictions than all baselines across the data sets. It is interesting to compare the accuracy between the single segment and sliding window baselines. For users where the single segment baseline is more accurate than the sliding window baseline (indicated by a lower normalized difference in Fig. 9), most of the predictive power is seen in the co-appearance structure in emails, e.g., there can be multiple "active" groups on any given day. Author C

and the synthetic user are two examples of such users—the importance of co-appearance information is apparent for the synthetic user in Figs. 4 and 5, where there is significant group overlap both in terms of membership and time. For users where the sliding window baseline is more accurate, most of the predictive power is seen in the appearance of email recipients over time. Authors A and B have better predictive performance with the sliding window baseline; these users focus on sending emails to a few select groups for sustained periods of time, and then focus on other groups as time progresses. The results show that the model we proposed in this paper can harness both temporal patterns and group structure for predictive purposes.

8 Discussion and conclusion

While the model proposed in this paper is a useful starting point for modeling data such as email histories, there are a wide variety of potential extensions and generalizations that are worth exploring. For example, the Poisson regression framework we employ is quite flexible, and one could use it to incorporate other exogenous covariates as well as detecting global segment boundaries that affect all groups and not just a single group. Furthermore, the real data often exhibits intermittent bursts of activity "embedded" within longer sequences of lower-level activity, suggesting that a model allowing temporal bursts (e.g., as in Kleinberg 2003), superposed on the segments, may be a useful avenue for further exploration.

In our model, we used a piecewise-constant framework for the group-dependent Poisson processes, where the emailing rates are constant between changepoints. While this is able to accurately reflect sudden changes in user behavior, it is clearly an approximation to real email activity. A natural extension to the model would be to augment the piecewise-constant framework to also incorporate gradual changes over time as well as trends.

There are also numerous opportunities to extend the modeling of groups. For example, in the present work we fix the number of clusters, K, but one could include a second non-parametric component to the group component of the model by allowing each email the opportunity to be sent to a newly created group of recipients. It would also be natural to allow groups to be related and dependent (e.g., via hierarchies) as well as to allow the group memberships (e.g., the independent Bernoulli probabilities) to change over time, e.g., as new people join a project and others leave.

In conclusion, we have presented a statistical model for exploring and analyzing egocentric email networks over time. This model can find interpretable groups of individuals by leveraging both co-appearance in individual emails as well as co-appearance during similar times of activity. We illustrated the exploratory aspects of our approach by fitting the model to data from multiple real email accounts and interpreting the composition of the learned groups and the parameters governing their prevalence over time. In addition, experiments indicated that the model yields improved predictive accuracy and coherence of groups over a variety of baselines. While the model in the paper was described in context of sending emails, it can be readily applied to broader types of multi-recipient directed communication data.

Acknowledgements This work was supported in part by a National Defense Science and Engineering Graduate Fellowship (Christopher DuBois, Nicholas Navaroli), a Google Research Award (Padhraic Smyth), and by the Office of Naval Research/Multidisciplinary University Research Initiative under grant number N00014-08-1-1015 (Padhraic Smyth, Christopher DuBois, Nicholas Navaroli).

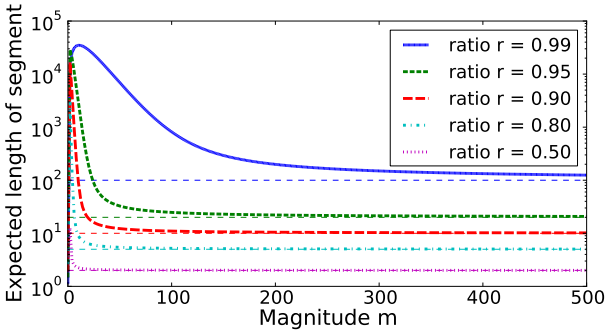


Fig. 10 Expected length of a segment under the prior for various settings of $m = \gamma + \kappa$ and $r = \frac{\gamma}{\gamma+\kappa}$. *Dashed lines* represent expected values for Geometric($\frac{\kappa}{\gamma+\kappa}$) distributions

Appendix A: Properties of the Dirichlet process prior

In this appendix we explore the significance of the Dirichlet Process hyperparameters γ and κ in Sect. 3.1, and the expected length of a segment in the HMM under this prior. For group k, the emailing rates $\{\lambda_{k,t} : 1 \leq t \leq T\}$ in the Poisson process depend on the HMM latent states $s_{k,t}$. The transition probabilities between latent states in the HMM depend on γ and κ. When a new state is transitioned to in the HMM, the probability of staying in the new state is the ratio $\frac{\gamma}{\gamma+\kappa}$. The magnitude $\gamma + \kappa$ represents the prior certainty of this ratio.

Setting values for γ and κ places a prior expectation on the length of a segment. Let v_x be the amount of time the HMM spends in segment x (where all $s_{k,t} = x$). The expected value for v_x can be derived to be

$$E[v_x|\gamma,\kappa] = \sum_{n=1}^{\infty} n P(v_k = n|\gamma,\kappa)$$

$$= \sum_{n=1}^{\infty} n \left(\prod_{i=0}^{n-1} \frac{i+\gamma}{i+\gamma+\kappa} \right) \frac{\kappa}{n+\gamma+\kappa},$$

assuming the time series is infinite. $P(v_k = n|\gamma,\kappa)$ is calculated by using (1). Figure 10 shows the expected value for several values of the ratio $\frac{\gamma}{\gamma+\kappa}$, each for a range of magnitude $\gamma + \kappa$. This expectation is estimated by truncating the above summation to $n = 1000000$.

When the magnitude of $\gamma + \kappa$ is large, V_t in (1) becomes negligible and the expected time in a segment approaches the expected value for a Geometric($\frac{\kappa}{\gamma+\kappa}$) distribution (as in a traditional HMM). When the magnitude is small, the expected value becomes much larger, since a single self-transition will make the probability of leaving that state approximately zero.

To illustrate the behavior of the prior under various hyperparameter settings, in Fig. 11 we show the posterior probability of a changepoint under different values of γ and κ for the synthetic email user in Sect. 5. Different configurations of magnitude $m = \gamma + \kappa$ and ratio $r = \frac{\gamma}{\gamma+\kappa}$ were considered. As the ratio becomes closer to 1, the prior is strong and favors large segments, regardless of the magnitude. This can be seen by the few peaks in changepoint probability when $r = 0.999$. As the ratio becomes smaller, the magnitude becomes increasingly important. When the magnitude is large, the a priori expected length of the segment becomes $\frac{r}{1-r}$. As a result, smaller segments are found. Evidence for this is seen in the significant probability of a changepoint across all days in the bottom-left plots in Fig. 11. When the magnitude is small, the prior favors large segments regardless of the ratio, resulting in fewer peaks in the posterior distributions of changepoints.

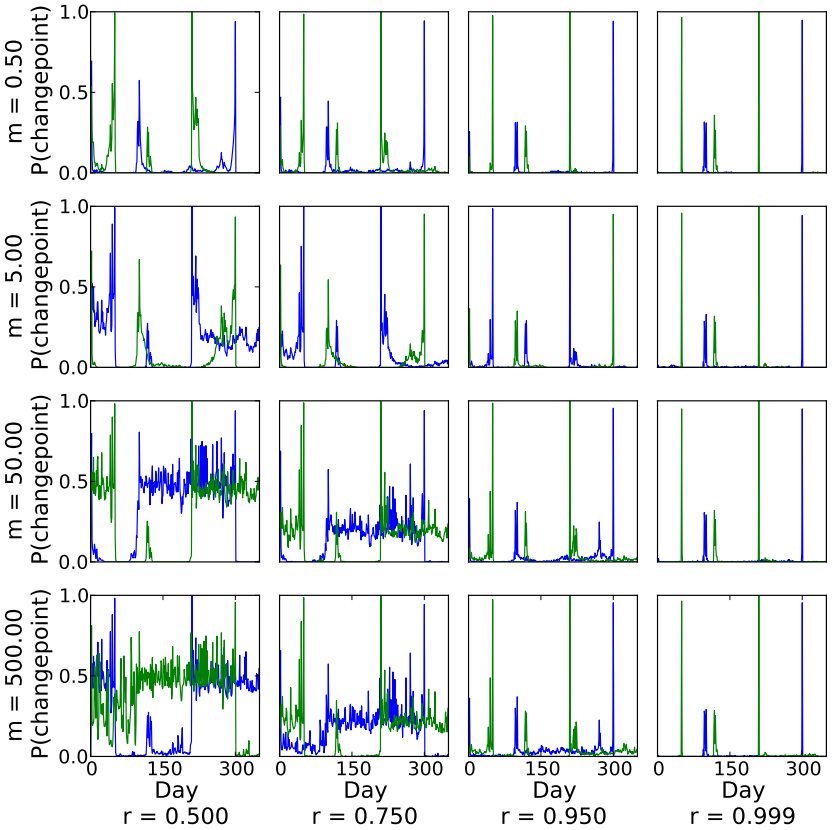


Fig. 11 Posterior probability of a changepoint for the synthetic email user in Sect. 5, using various values of $m = \gamma + \kappa$ and $r = \frac{\gamma}{\gamma+\kappa}$

Appendix B: Derivations of sampling equations

In this appendix we briefly derive the sampling equations used for sampling the model parameters, as described in Sect. 4. Throughout this appendix, we use the notation $A \setminus B$ to represent the elements in set A, not including those in set B.

B.1 Sampling equation for latent group assignments, Z

The Gibbs sampler samples $z_{t,n}$, the group assignment for email n on day t, by first calculating $P(z_{t,n} = k|\Theta \setminus z_{t,n})$ for all k, then sampling from this discrete distribution. This distribution is written as

$$P(z_{t,n}|\Theta \setminus z_{t,n}) \propto P\big(z, y|\Theta \setminus \{z, y\}\big)$$
$$= P(z|\alpha, \beta, s) \int P(y|z, \phi) P\big(\phi|\alpha^{(z)}, \beta^{(z)}\big) d\phi.$$

Expanding the term $P(y|z,\phi)$, we get

$$P(y|z,\phi) = \prod_{t'=1}^{T}\prod_{n'=1}^{N_{t'}}\prod_{r=1}^{R}\phi_{z_{t',n',r}}^{y_{t',n',r}}(1-\phi_{z_{t',n',r}})^{1-y_{t',n',r}}$$
$$= \prod_{k=1}^{K}\prod_{r=1}^{R}\phi_{k,r}^{c_{1,k,r}}(1-\phi_{k,r})^{c_{0,k,r}},$$

where $c_{1,k,r}$ is the number of emails sent to group k that recipient r appears in, and $c_{0,k,r}$ is the number of emails sent to group k that recipient r does not appear in. Combining this with the $P(\phi|\alpha^{(z)},\beta^{(z)})$ term and moving the products outside the integral, we get

$$P(z_{t,n}|\Theta\setminus z_{t,n}) \propto P(z_{t,n}|\alpha,\beta,s)\prod_{k=1}^{K}\prod_{r=1}^{R}\int \phi_{k,r}^{c_{1,k,r}+\alpha^{(z)}-1}(1-\phi_{k,r})^{c_{0,k,r}+\beta^{(z)}-1}d\phi_{k,r}$$
$$\propto P(z_{t,n}|\alpha,\beta,s)\prod_{k=1}^{K}\prod_{r=1}^{R}\frac{\Gamma(c_{1,k,r}+\alpha^{(z)})\Gamma(c_{0,k,r}+\beta^{(z)})}{\Gamma(c_{1,k,r}+\alpha^{(z)}+c_{0,k,r}+\beta^{(z)})}$$

since the integral has the form of an unnormalized Beta$(c_{1,k,r}+\alpha^{(z)}, c_{0,k,r}+\beta^{(z)})$ distribution. Consider the above fraction for the product term $k=z_{t,n}$. We define $c_{1,k,r}^{-(t,n)}$ to be the same as $c_{1,k,r}$, however email n on day t is ignored. Similarly for $c_{0,k,r}^{-(t,n)}$. The above fraction for $k=z_{t,n}$ can be rewritten as

$$\prod_{r=1}^{R}\frac{\Gamma(c_{1,z_{t,n},r}^{-(t,n)}+y_{t,n,r}+\alpha^{(z)})\Gamma(c_{0,z_{t,n},r}^{-(t,n)}+(1-y_{t,n,r})+\beta^{(z)})}{\Gamma(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)}+c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)}+1)}.$$

Using the property $\Gamma(x+1)=x\Gamma(x)$, we have

$$\prod_{r=1}^{R}\left(\frac{\Gamma(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)})\Gamma(c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)})}{\Gamma(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)}+c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)})}\right)\left(\frac{(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)})^{y_{t,n,r}}(c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)})^{(1-y_{t,n,r})}}{c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)}+c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)}}\right).$$

The left fraction can now be absorbed in the product over different groups, as $c_{\{0,1\},k',r}^{-(t,n)}$ is equivalent to $c_{\{0,1\},k',r}$ for $k'\neq z_{t,n}$.

$$P(z_{t,n}|\Theta\setminus z_{t,n}) \propto P(z_{t,n}|\alpha,\beta,s)\left(\prod_{k=1}^{K}\prod_{r=1}^{R}\frac{\Gamma(c_{1,k,r}^{-(t,n)}+\alpha^{(z)})\Gamma(c_{0,k,r}^{-(t,n)}+\beta^{(z)})}{\Gamma(c_{1,k,r}^{-(t,n)}+\alpha^{(z)}+c_{0,k,r}^{-(t,n)}+\beta^{(z)})}\right)$$
$$\times\left(\prod_{r=1}^{R}\frac{(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)})^{y_{t,n,r}}(c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)})^{(1-y_{t,n,r})}}{c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)}+c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)}}\right)$$
$$\propto \lambda_{z_{t,n},t}\left(\prod_{r=1}^{R}\frac{(c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)})^{y_{t,n,r}}(c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)})^{(1-y_{t,n,r})}}{c_{1,z_{t,n},r}^{-(t,n)}+\alpha^{(z)}+c_{0,z_{t,n},r}^{-(t,n)}+\beta^{(z)}}\right)$$

B.2 Sampling equation for day-of-week regression parameters, α

In order to sample α_w, the regression parameter for day-of-week w, we need the unnormalized log conditional distribution. The distribution can be written as

$$\begin{aligned} P(\alpha_w|\Theta \setminus \alpha_w) &\propto P\big(\alpha_w, N|\Theta \setminus \{\alpha_w, N\}\big) \\ &\propto P\big(\alpha_w|\mu, \sigma^2\big) \prod_{t:d(t)=w} P(N_t|\alpha, \beta, s), \end{aligned}$$

where $d(t)$ is the day-of-week for day t. From Sect. 3.1, we have

$$P(N_t|\alpha, \beta, s) = \text{Poisson}\left(\sum_{k=1}^{K} \lambda_{k,t}\right) = \text{Poisson}\left(e^{\alpha_w} \sum_{k=1}^{K} e^{\beta_{k,s_{k,t}}}\right).$$

Taking the log of this unnormalized conditional distribution, we get

$$\begin{aligned} \log P(\alpha_w|\Theta \setminus \alpha_w) &\propto -\frac{(\alpha_w - \mu)^2}{2\sigma^2} \\ &\quad + \sum_{t:d(t)=w} \left[N_t \left(\alpha_w + \log\left(\sum_{k=1}^{K} e^{\beta_{k,s_{k,t}}} \right) \right) - e^{\alpha_w} \left(\sum_{k=1}^{K} e^{\beta_{k,s_{k,t}}} \right) \right] \\ &\propto -\frac{(\alpha_w - \mu)^2}{2\sigma^2} + \alpha_w \sum_{t:d(t)=w} N_t - e^{\alpha_w} \left(\sum_{t:d(t)=w} \sum_{k=1}^{K} e^{\beta_{k,s_{k,t}}} \right). \end{aligned}$$

B.3 Sampling equation for group segment regression parameters, β

As with α, we sample using the unnormalized log conditional distribution. This distribution for $\beta_{k,m}$, the regression coefficient for group k while in segment m, is

$$\begin{aligned} P(\beta_{k,m}|\Theta \setminus \beta_{k,m}) &\propto P\big(\beta_{k,m}, N, z|\Theta \setminus \{\beta_{k,m}, N, z\}\big) \\ &\propto P\big(\beta_{k,m}|\mu, \sigma^2\big) \prod_{t:s_{k,t}=m} P(N_t|\alpha, \beta, s) \prod_{t:s_{k,t}=m} \prod_{n=1}^{N_t} P(z_{t,n}|\alpha, \beta, s). \end{aligned}$$

Using (4) for $P(z_{t,n}|\alpha, \beta, s)$, substituting λ for $\{\alpha, \beta, s\}$, and taking the log of this unnormalized conditional distribution, we get

$$\begin{aligned} \log P(\beta_{k,m}|\Theta \setminus \beta_{k,m}) &\propto -\frac{(\beta_{k,m} - \mu)^2}{2\sigma^2} + \sum_{t:s_{k,t}=m} \left[N_t \log\left(\sum_{k'=1}^{K} \lambda_{k',t} \right) - \sum_{k'=1}^{K} \lambda_{k',t} \right] \\ &\quad + \sum_{t:s_{k,t}=m} \left(\sum_{n=1}^{N_t} \log(\lambda_{z_{t,n},t}) - N_t \log\left(\sum_{k'=1}^{K} \lambda_{k',t} \right) \right) \\ &\propto -\frac{(\beta_{k,m} - \mu)^2}{2\sigma^2} - \sum_{t:s_{k,t}=m} \lambda_{k,t} + \sum_{t:s_{k,t}=m} \sum_{n=1}^{N_t} \log(\lambda_{z_{t,n},t}). \end{aligned}$$

Using (2), we get

$$\log P(\beta_{k,m}|\Theta \setminus \beta_{k,m}) \propto -\frac{(\beta_{k,m}-\mu)^2}{2\sigma^2} - \sum_{t:s_{k,t}=m} e^{\alpha_{d(t)}} e^{\beta_{k,m}} + \big|\{t,n : s_{k,t}=m, z_{t,n}=k\}\big|\beta_{k,m}.$$

B.4 Sampling equation for HMM latent states, s

Because of the HMM restrictions described in Sect. 3.1, the Gibbs sampler only samples the HMM latent state $s_{k,t}$ if $s_{k,t-1} \neq s_{k,t}$ or $s_{k,t} \neq s_{k,t+1}$. When a new value for $s_{k,t}$ is sampled, its possible values are $s_{k,t-1}$, $s_{k,t+1}$, or a brand new state. Thus, the sampler samples from the discrete distribution defined by the probabilities of these outcomes. The unnormalized conditional probability for $s_{k,t}$ is

$$\begin{aligned} P(s_{k,t}|\Theta \setminus s_{k,t}) &\propto P\big(s, z, N|\Theta \setminus \{s, z, N\}\big) \\ &\propto P(N_t|\alpha,\beta,s) \prod_{n=1}^{N_t} P(z_{t,n}|\alpha,\beta,s) \prod_{t'=t}^{g(t)} P\big(s_{k,t'} \mid S\big(t'\big), \gamma, \kappa\big), \end{aligned}$$

where $S(t') = \{s_{k,u} : u < t'\}$ and $g(t) = \min(\{t' : t' \geq t+1, s_{k,t'} \neq s_{k,t+1}\})$, the time of the start of the segment following the one $s_{k,t+1}$ is in. Usually, the last product is over $t' = \{t, t+1\}$ because of the Markov dependence between latent states in a HMM. However, the transition matrices for the HMMs are integrated out, introducing the additional dependencies (as seen in the V_t variable in (1)). Using (1), the last product can be written as

$$\prod_{t'=t}^{g(t)} \frac{(V_{t'}+\gamma)^{\delta(s_{k,t'}=s_{k,t'-1})} \kappa^{\delta(s_{k,t'}\neq s_{k,t'-1})}}{V_{t'}+\gamma+\kappa}.$$

The new value of $s_{k,t}$ is sampled by first calculating the unnormalized probabilities above for each value of $s_{k,t}$ (V_t changes based on $s_{k,t}$), then normalizing and sampling from the corresponding discrete distribution.

References

Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*, 1981–2014.

Alison Bryant, J., Sanders-Jackson, A., & Smallwood, A. (2006). IMing, text messaging, and adolescent social networks. *Journal of Computer-Mediated Communication*, *11*(2), 577–592.

Beal, M., Ghahramani, Z., & Rasmussen, C. (2002). The infinite hidden Markov model. In *Advances in neural information processing systems* (Vol. 14, pp. 577–584). Cambridge: MIT Press.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). New York: ACM.

Butts, C. (2008). A relational event framework for social action. *Sociological Methodology*, *38*(1), 155–200.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*(2), 221–241.

Choi, D., Wolfe, P., & Airoldi, E. (2012). Stochastic blockmodels with growing number of classes. *Biometrika*, *99*(2), 273–284.

de Nooy, W. (2011). Networks of action and events over time. a multilevel discrete-time event history model for longitudinal network data. *Social Networks*, *33*(1), 31–40.

Diesner, J., Frantz, T., & Carley, K. (2005). Communication networks from the Enron email corpus: It's always about the people. Enron is no different. *Computational and Mathematical Organization Theory*, *11*(3), 201–228.

Dredze, M., Schilit, B., & Norvig, P. (2009a). Suggesting email view filters for triage and search. In *Proceedings of the international joint conference on artificial intelligence* (Vol. 9, pp. 1414–1419).

Dredze, M., Wallach, H., Puller, D., Brooks, T., Carroll, J., Magarick, J., Blitzer, J., & Pereira, F. (2009b). Intelligent email: aiding users with AI. In *Proceedings of the twenty-third AAAI conference on artificial intelligence* (pp. 1524–1527). Menlo Park: AAAI Press.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, *16*, 203–213.

Fisher, D. (2005). Using egocentric networks to understand communication. *IEEE Internet Computing*, *9*(5), 20–28.

Fisher, D., Brush, A., Gleave, E., & Smith, M. (2006). Revisiting Whittaker & Sidner's email overload ten years later. In *Proceedings of the 2006 20th anniversary conference on computer supported cooperative work* (pp. 309–312). New York: ACM.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, *5*(2A), 1020–1056.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *J. Comput.-Med. Commun.*, *3*(1).

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*(4), 373–397.

Klimt, B., & Yang, Y. (2004). The Enron corpus: a new dataset for email classification research. In *Proceedings of the European conference on machine learning* (pp. 217–226). Berlin: Springer.

Koren, Y., Liberty, E., & Sandler R, M. Y. (2011). Automatically tagging email by leveraging other users' folders. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 913–921). New York: ACM.

MacLean, D., Hangal, S., Teh, S., Lam, M., & Heer, J. (2011). Groups without tears: mining social topologies from email. In *Proceedings of the 16th international conference on intelligent user interfaces* (pp. 83–92). New York: ACM.

Malmgren, R., Stouffer, D., Motter, A., & Amaral, L. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 18,153–18,158.

Malmgren, R., Hofman, J., Amaral, L., & Watts, D. (2009). Characterizing individual communication patterns. In *Proceedings of the 15th ACM SIGKDD conference* (pp. 607–616). New York: ACM.

McAuley, J., & Leskovec, J. (2012). Learning to discover social circles in ego networks. In *Advances in neural information processing systems* (pp. 548–556).

Navaroli, N., DuBois, C., & Smyth, P. (2012). Statistical models for exploring individual email communication behavior. In *Proceedings of the 4th Asian conference on machine learning (ACML 2012), JMLR workshop and conference proceedings* (Vol. 25, pp. 317–332).

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Neal, R. (2003). Slice sampling. *The Annals of Statistics*, *31*, 705–741.

Ross, S. M. (2007). *Introduction to probability models* (9th ed.). San Diego: Academic Press.

Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., & Merom, R. (2010). Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD international conference* (pp. 233–242). New York: ACM.

Scott, S. (2004). A Bayesian paradigm for designing intrusion detection systems. *Computational Statistics & Data Analysis*, *45*, 69–83.

Scott, S., & Smyth, P. (2003). The Markov modulated Poisson process and Markov Poisson cascade with applications to Web traffic data. *Bayesian Statistics*, *7*, 671–680.

Tantipathananandh, C., Berger-Wolf, T., & Kempe, D. (2007). A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference* (pp. 717–726). New York: ACM.

Wainer, J., Dabbish, L., & Kraut, R. (2011). Should I open this email?: inbox-level cues, curiosity and attention to email. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 3439–3448). New York: ACM.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In *The 23rd conference on uncertainty in artificial intelligence* (pp. 579–586).

Whittaker, S., Matthews, T., Cerruti, J., Badenes, H., & Tang, J. (2011). Am I wasting my time organizing email?: a study of email refinding. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 3449–3458). New York: ACM.

Wolfram, S. (2012). The personal analytics of my life. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

Xu, K., Kliger, M., & Hero, A. (2011). Tracking communities in dynamic social networks. In *Proceedings of the 4th international conference on social computing, behavioral-cultural modeling and prediction* (pp. 219–226). Berlin: Springer.

Zenk, L., Stadtfeld, C., & Windhager, F. (2010). How to analyze dynamic network patterns of high performing teams. *Procedia Social and Behavioral Sciences*, *2*(4), 6418–6422.