JMLR: Workshop and Conference Proceedings 25:317–332, 2012 Asian Conference on Machine Learning

Statistical Models for Exploring Individual Email Communication Behavior

Nicholas Navaroli NNAVAROL@ICS.UCI.EDU
Department of Computer Science
University of California, Irvine

Christopher DuBois DUBOISC@ICS.UCI.EDU
Department of Statistics
University of California, Irvine

Padhraic Smyth SMYTH@ICS.UCI.EDU
Department of Computer Science
University of California, Irvine

Editor: Steven C.H. Hoi and Wray Buntine

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, 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. In particular, in this paper we focus on analyzing egocentric network data over time (and more specifically, email histories) consisting of time series of counts of communication events between an ego and his or her alters. Our goal is to develop a statistical model that can summarize the major characteristics of such data: who does the ego communicate with? at what rates? and how do these patterns change over time?

As an example, Figure 1 shows the weekly email communication patterns from several years of email history from one of the authors of this paper. In the right plot the x-axis represents time and the y-axis represents different recipients, with dots representing which recipients receive an email on which 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,

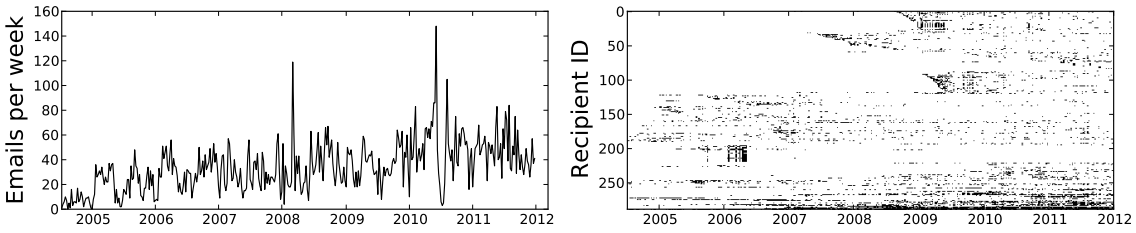


Figure 1: Personal email communication data for one of the authors. Left: Total number of emails sent per week. Right: Points indicate an email sent to a particular individual (y-axis) at a particular week (x-axis).

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. Our goal is to develop an unsupervised learning approach based on a statistical model 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 such a model. Individual users can use this type of model to better understand their digital behavior over time. An example is the recent publication by Stephen Wolfram on his blog of detailed visualizations and commentary on 13 years worth of his personal email data (Wolfram, 2012). Companies providing communication services (such as email or social network channels) can analyze their users' usage of such services and potentially enhance the user experience from a better understanding of user behavior over time. A simple example of this type 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. In the social sciences, there is increasing interest in analyzing digital human communication data to inform social theorists about human behavior in the context of modern communication media (e.g., Butts, 2008).

In this paper we focus specifically on modeling daily 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 simpler than modeling both senders and receivers. Of the two, sent emails are potentially of more interest in informing us about the individual since they are the result of specific actions by the individual, while received emails 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 or latent Dirichlet analysis 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 novel 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 Section 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 Section 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 compares the predictive accuracy of the proposed model to alternative baseline approaches. Section 8 contains discussion and conclusions.

2. Related Work

Prior work of relevance to our proposed approach can be broadly categorized into 3 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 communication data over time 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 signif-

icant 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 advances in using non-parametric Bayesian methods for other types of human communication, such as detecting changes in speaker from audio recordings of meetings (Fox et al., 2011). These methods use a hidden Markov model 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 the topic 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 a particular time. We take an approach akin to mixed membership models (Airoldi et al., 2008; Choi et al., 2012), allowing individuals to be members of multiple groups. In particular, we jointly model both the group memberships and the rate of events involving a particular group, which contrasts to methods whose sole focus is the progression of latent group memberships at discrete timesteps. 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 Section 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 Section 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 ($N_t = \sum_{k=1}^{K} N_{k,t} \sim \text{Poisson}(\sum_{k=1}^{K} \lambda_{k,t})$).

The Poisson Process for a Single Group

We begin by describing (in the remainder of this subsection) the model for time-varying communication rates for a single group, deferring discussion of how we learn the groups themselves to Section 3.2. We model a user's email rate to group k, $\{\lambda_{k,t} : 1 \leq t \leq T\}$, using a hidden Markov model. 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 betwen 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:

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

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 long the HMM has been in state $s_{k,t-1}$ up to time t.

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

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.

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). In the regression content, $X_{k,t}$ is a binary vector indicating the latent state of the HMM on day t, and $\theta = [\beta_{k,1}, \beta_{k,2}, \cdots, \beta_{k,M_k}]^T$, 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, we can include *day-of-week* effects:

$$\log \lambda_{k,t} = \beta_{k,s_{k,t}} + \alpha_{w_t} \tag{1}$$

where $w_t \in W$ represents different days of the week. We can use $W = \{0, 1\}$ to represent weekdays and weekends, or $W = \{0, \cdots, 6\}$ to represent each day of the week individually (we use the latter for all results in this paper). The corresponding α terms capture routine changes in behavior on a weekly basis. For example, if a user only sends emails on the weekdays the α_{weekend} term would have a large negative value, making $\lambda_{k,t} \approx 0$ on weekends.

3.2. Modeling Recipient Groups

We now discuss how to model the K different groups that the user interacts with, where a group is defined as a distribution over R possible recipients for an email. The goal is to have different groups model different sets of people that share a common characteristic with the user (such as common familial relationships, organizational relationships, common interests, and so forth).

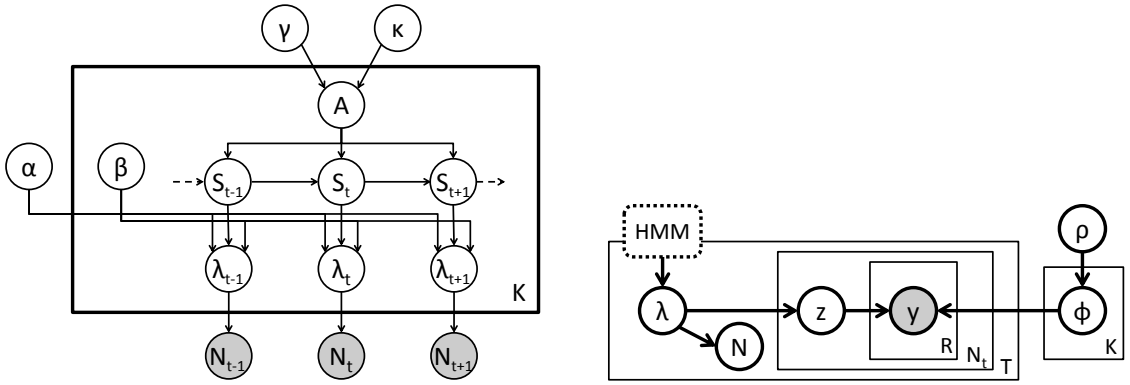
We assume each email is sent to one of the K latent groups. Let $z_{t,n}$ represent the latent group that email n on day t was sent to. Given the latent group, the recipients of the email are chosen from a vector of probabilities $\phi_{z_{t,n}}$ of length R. The generative model for the recipients given the latent group is

$$\phi_k | \rho \sim \text{Dirichlet}(\rho) \qquad\qquad y_{t,n} | \phi, z_{t,n} \sim \text{Multinomial}(m_{t,n}, \phi_{z_{t,n}})$$

where $y_{t,n} \in [0, 1]^R$ is a binary vector indicating which recipients are in email n on day t, $m_{t,n}$ is the number of recipients in the email, and ρ is a vector of length R corresponding to the parameters of the Dirichlet prior that helps to smooth the probability estimates. In the results below we set all $\rho_r = 1$. Note that a multinomial model allows for the unrealistic possibility of the same individual receiving the same email more than once. However, conditioned on the observed data, where each recipient is included in the recipient list for an email only once, the multinomial likelihood is nonetheless a useful way to compute the probability of a such a list for each group k. Alternatives to the multinomial that could be used in this context include a conditional independence Bernoulli model or a multivariate hypergeometric distribution. We chose to use the multinomial model for convenience.

Determining Latent Group Probabilities

The modeling of the latent group indicator variables $z_{t,n}$ is a key aspect of the model; the distribution of $z_{t,n}$ connects the K separate HMMs from Section 3.1 and the generative model of email recipients in Section 3.2. The multinomial 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 $\{\lambda_{k,t} : 1 \leq t \leq T\}$ is a Poisson process and the time between



(*a*) Graphical representation of the K HMMs associated with the K groups. A represents HMM transition matrices.

(*b*) Graphical representation for the K groups. The "HMM" node is a compact representation of Figure 2(*a*).

Figure 2: The graphical representation of the model.

emails for a particular group follows an exponential distribution, it is straightforward to show (using Equation 1) that the probability of the next email being sent to group k can be written as:

$$P(z_{t,n} = k | \{\lambda_{k,t} : 1 \leq k \leq K\}) = \frac{\lambda_{k,t}}{\sum_{k'=1}^{K} \lambda_{k',t}} = \frac{e^{\alpha_{w_t}} e^{\beta_{k,s_{k,t}}}}{\sum_{k'=1}^{K} e^{\alpha_{w_t}} e^{\beta_{k',s_{k',t}}}} = \frac{e^{\beta_{k,s_{k,t}}}}{\sum_{k'=1}^{K} e^{\beta_{k',s_{k',t}}}}$$

Our approach is similar in some respects to the dynamic topic model approach of Blei and Lafferty (2006). Instead of associating each latent topic variable with a word token, in our model each latent variable is associated with the set of recipients in an email. In addition, to model the changes in group behavior over time, we use a discrete-state Markov process to explicitly model changepoints in the user's behavior, instead of an autoregressive approach.

Figure 2 shows the graphical model for representing the HMM part of the model in 2(*a*) and the group aspect of the model in 2(*b*). In the interests of interpretability, all the HMM variables (latent states, transition matrix, regression parameters, and Dirichlet process priors) are combined into a single supernode in Figure 2(*b*). Note that, since the rate of sending emails λ is a deterministic function of the regression parameters α and β, λ can be removed from the graphical model. We keep λ in the graphical model for clarity.

4. The MCMC Inference Algorithm

We use Markov chain Monte Carlo techniques to learn the parameters of our model from observed data. 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 Figures 2(*a*) and 2(*b*), 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. To keep the notation simple, variables without subscripts denote the set of all variables that can be indexed by it, e.g., $\lambda = \{\lambda_{k,t} : 1 \leq k \leq K, 1 \leq t \leq T\}$.

4.1. Sampling the latent groups

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

$$P(z_{t,n} = k|\cdot) \propto P(z|\lambda) \int P(y|z,\phi)P(\phi|\rho)d\phi \quad \propto \frac{\lambda_{k,t}}{\sum_{k'=1}^{K} \lambda_{k',t}} \frac{\prod_{r=1}^{R}(c_{k,r}^{-(t,n)} + \rho_r)}{\prod_{i=0}^{m_{t,n}-1}(i + \sum_{r'=1}^{R}(c_{k,r'}^{-(t,n)} + \rho_{r'}))}$$

where $c_{k,r}^{-(t,n)} = \sum_{t' \neq t} \sum_{n' \neq n} y_{t',n',r} \delta(z_{t',n'} = k)$ is the number of times recipient r was present in an email sent to group k, ignoring email n on day t, and $m_{t,n}$ is the number of recipients for email n on day t. The derivation of this conditional distribution is similar to that of the standard collapsed Gibbs sampling equations for LDA.

4.2. Sampling the regression parameters

We place a non-conjugate Normal(μ, σ^2) prior on each of the regression parameters. We can sample from this conditional distribution, using the unnormalized log distribution, via a technique known as slice sampling (Neal, 2003). The conditional distributions for sampling the regression parameters $\{\alpha_w\}$ and $\{\beta_{k,m}\}$, are

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

where $g_{k,m} = \sum_{t=1}^{T} \delta(s_{k,t} = m) \sum_{n=1}^{N_t} \delta(z_{t,n} = k)$, the number of times an email was sent to group k when that group was in segment m. Note that updating both α and β automatically updates the emailing rates λ.

4.3. Sampling the HMM hyperparameters

While it is possible to place priors over the Dirichlet process hyperparameters γ and κ, we instead define priors over their ratio $r = \frac{\gamma}{\gamma+\kappa}$ and magnitude $m = \gamma + \kappa$. The ratio represents the probability of staying in a newly visited state, and the magnitude represents the strength of the prior. As priors we use $m \sim \text{Gamma}(k_g, \theta_g)$ and $r \sim \text{Beta}(\alpha_b, \beta_b)$. As with the regression parameters, these priors are non-conjugate, so we use slice sampling over the conditional unnormalized log 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:

$$P(m|\cdot) \propto P(m|k_g,\theta_g)P(s|\gamma,\kappa) \qquad P(r|\cdot) \propto P(r|\alpha_b,\beta_b)P(s|\gamma,\kappa)$$

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

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

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 represents 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 Figure 3 show the membership probabilities of the two groups; each group has 3 exclusive recipients, with the remaining 4 recipients seen in both groups. The top-left plot in Figure 4 shows the values of the β regression parameters, with each group being dominant during different periods of time. Group 0 has two changepoints on days 100 and 300, and group 1 has three changepoints on days 50, 120, and 210. The values for the α regression parameters were set so that the email rates on weekends are 60% of the rates during the weekdays.

Given the parameters of the model, emails are simulated by first simulating the total number of emails sent for each day; $N_t \sim \text{Poisson}(\sum_{k=1}^{K} \lambda_{k,t})$. The bottom-right plot in Figure 4 shows the sampled values for N_t. For each of the N_t emails, we simulate which group the email was sent to; $z_{t,n} \sim \text{Multinomial}(\hat{\lambda}_t)$, where $\hat{\lambda}_t$ are the normalized $\lambda_{k,t}$, for $1 \leq k \leq K$. Each email is equally likely to contain one or two recipients.

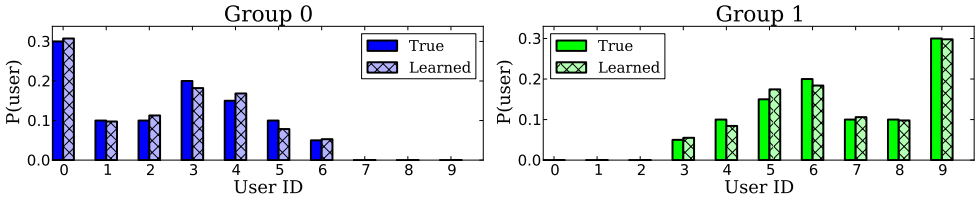


Figure 3: The true and learned membership probabilities for the two different groups.

To learn the parameters of the model, we iteratively sample the parameters of our model as described in Section 4. 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 1 day. The regression and Dirichlet process parameters are initialized to a sample from their prior distributions. The regression parameters have a Normal$(0, 1)$ prior, the magnitude $\gamma + \kappa$ has a Gamma$(0.5, 20)$ prior, and the ratio $\frac{\gamma}{\gamma+\kappa}$ has a Beta$(10, 1)$ prior. Lastly, the group parameters ϕ_k have Dirichlet(ρ) priors, where ρ is a vector of ones of length R.

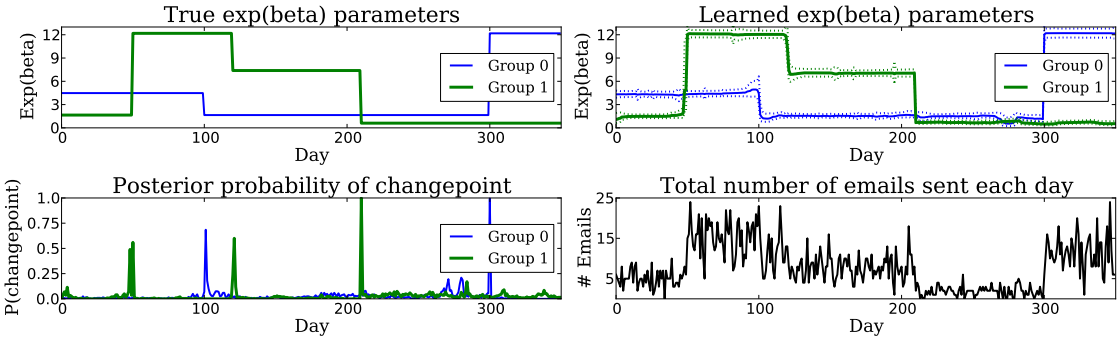


Figure 4: The β regression parameters used to generate the email traffic. The model is able to learn the correct values of β and locations of changepoints.

The latent group variables $z_{t,n}$ are initialized uniformly at random between the two groups. After collecting 2100 samples, the first 100 are discarded as burnin, and every tenth sample after that is kept (for a total of 200 kept samples).

Figures 3 and 4 show the learned parameters, alongside the true parameters. The cross-hatched bars in Figure 3 show the learned groups. Membership probabilities are recovered using the set of latent z variables from the sample that produced the largest log-likelihood (maximum a posteriori (MAP) estimate). The top-right plot in Figure 4 shows the average β regression parameters across the 200 samples, with dashed lines showing one standard deviation. 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 also learned the correct α parameters (results not shown). The bottom-left plot in Figure 4 shows the posterior probability of changepoints for the two groups. The posterior probability of a changepoint on day t for group k is the fraction of samples where $s_{k,t} \neq s_{k,t+1}$.

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 on 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. Email addresses of the recipients were then converted to anonymous user ids. For all of the email data sets we filtered out recipients that receive less than 10 emails.

For each data set, we learn the parameters of the model using $K = 50$ groups. We experimented with values of K ranging from 10 to 100 and generally found that K in the range

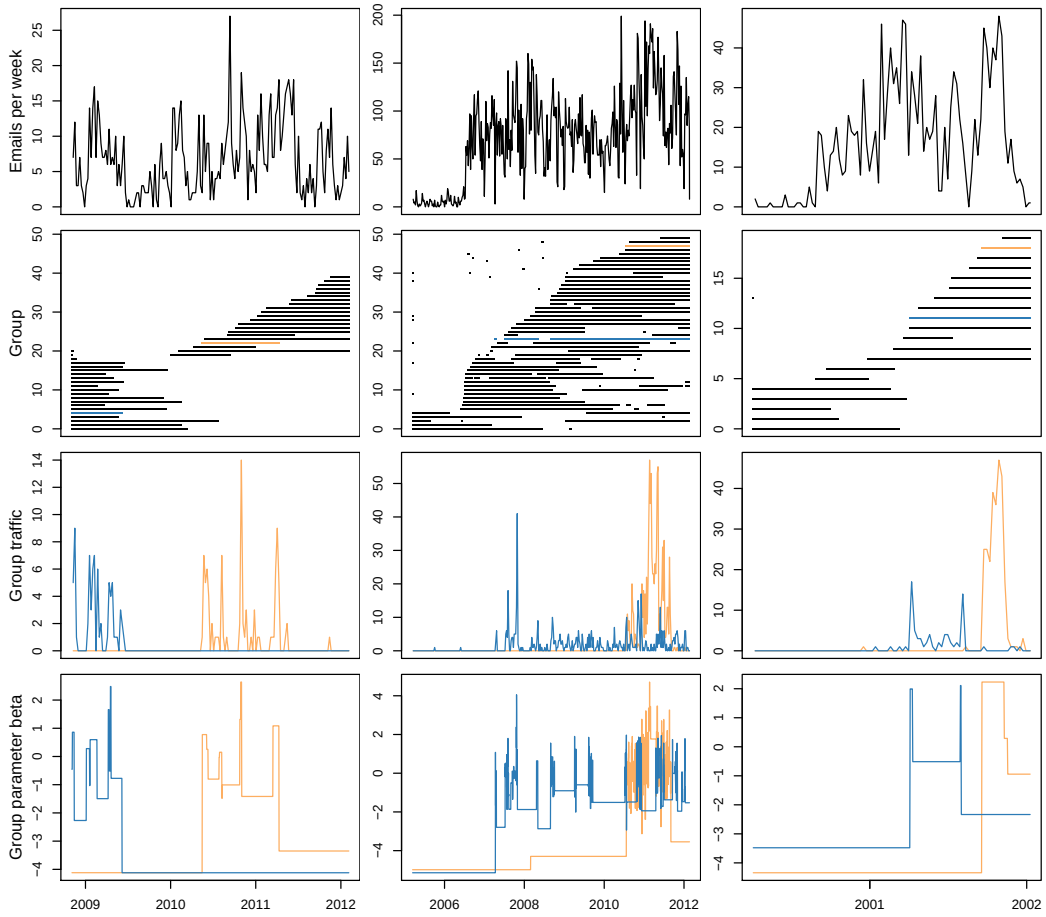


Figure 5: Exploratory analysis from fitting the model to real-world email data. Each column corresponds to the email history of a specific individual. First row: Observed weekly email volume. Second row: greater than average β values for each of the $K = 50$ learned groups. Third row: Number of emails per week assigned to two chosen groups, highlighted throughout as blue and orange. Fourth row: The learned parameters β for the chosen groups. See text for details.

of 20 to 50 produced interpretable results on real-world data sets. An obvious extension to the model would be to learn K, e.g., using Bayesian non-parameteric techniques. We use the same initialization of model parameters and configurations of the hyperparameters as the synthetic data in the previous section. The sampler does not discard any samples for burnin and every 200$^{\text{th}}$ sample is kept until 500 samples are collected (a total of 100,000 samples are produced).

Figure 5 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, using the same methodology we used for the simulated data. 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 is one of the most active users from the Enron corpus (Klimt and Yang, 2004).

The left column of Figure 5 shows the learned parameters of the model for one of the authors of this paper. There is clearly a major changepoint in this user's behavior, around the middle of 2009, when the user transitioned between institutions. This large-scale change shows that, for this user, 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, one from each university. All of the emails in the blue group were directed to people who held administrative positions at the pre-2009 institution. The emails in the orange group are sent to members in a specific research project, with spikes in activity corresponding to different deadlines.

The center column of Figure 5 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 specific learned groups. The blue group corresponds to a project where the author is a principal investigator for a large proposal and the recipients are 6 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 activity of the orange group ramps up in mid 2010 as the author took on organizational duties for a relatively large conference, followed by roughly 15 months of relatively high activity until the actual conference in summer 2011. The multinomial distribution for this group focused over 95% of the probability mass on about 15 recipients, all of whom were involved in program organization for the conference.

The third column in Figure 5 illustrates results for an active user from the Enron corpus. This user's email activity ramped up in early 2001 and there appears to have been a significant change in recipient groups around the same time—several of the groups (horizontal bars) in the second row end and several new ones begin.

We found similar interesting patterns for the other author of the paper and for other Enron users. As with other latent variable models such as LDA, while most groups were focused on a small subset of the recipients and were active during specific intervals of time, not all of the learned groups were completely intuitive or interpretable. For example, a few "singleton" groups were learned for some users, consisting of just a few emails sent to a

single person. This is probably a result of the number of groups being too high for this user and in effect the model is overfitting.

7. Experiments on Predictive Performance

In this section we measure the predictive performance of our model when recipients are removed uniformly at random from emails in the training data. As an example, suppose an email was sent to recipients A, B, C and we remove C in the training data. We then test the model by computing the conditional probability of C as a recipient given that A and B were observed. Let $y_{t,n}^{\text{obs}}$ be the observed recipients for email n on day t, and $y_{t,n}^{\text{miss}}$ be the missing recipients, such that $y_{t,n} = y_{t,n}^{\text{obs}} + y_{t,n}^{\text{miss}}$ (these are binary vectors). If all original recipients of an email are removed, that email is removed completely from the observed data set (decreasing the value of N_t for that day). The model is then trained on the remaining data, ignoring the missing recipients[2]. The predictive performance of the model (and baselines) on missing data is evaluated using the test log-likelihood:

$$\begin{aligned} LL_{\text{test}} &= \sum_{t=1}^{T}\sum_{n=1}^{N_t} \log\left(\sum_{k=1}^{K} P(z_{t,n}=k|\lambda,\phi,y_{t,n}^{\text{obs}})P(y_{t,n}^{\text{miss}}|\phi_k,y_{t,n}^{\text{obs}})\right) \\ &= \sum_{t=1}^{T}\sum_{n=1}^{N_t} \log\left(\sum_{k=1}^{K}\left[\frac{P(z_{t,n}=k|\lambda,\phi)P(y_{t,n}^{\text{obs}}|\phi_k)}{\sum_{k'=1}^{K} P(z_{t,n}=k'|\lambda,\phi)P(y_{t,n}^{\text{obs}}|\phi_{k'})}\right] P(y_{t,n}^{\text{miss}}|\phi_k)\right) \end{aligned}$$

In our experiments below we generated 10 training and test data sets in this manner, randomly putting 20% of recipients in the test set each time, and computing the average log-likelihood across the 10 test sets. For each training data set, the parameters of our model were learned by collecting 750 MCMC samples. The first 250 were discarded for burn-in, and every fifth sample was kept after that, leaving a total of 100 samples. The ϕ_k above were estimated from the latent group variables $\{z_{t,n}\}$ in the sample that produced the largest log-likelihood (MAP estimate), and the group rates $\{\lambda_{k,t}\}$ were estimated by taking the average value across the 100 samples. For each data set, the model was trained to learn $K = 50$ groups (for synthetic data, $K = 2$).

We compare below the predictive power of our model with 4 baseline approaches. *Uniform* is a uniform distribution over all possible recipients. *Single multinomial* corresponds to a maximum likelihood of a multinomial model over possible recipients. The *sliding window/no groups* model is similar to the single multinomial model, except that the multinomial is based on local time-windows, allowing the multinomial to adapt to changes over time in recipient likelihood. We evaluated different sized windows up to 2 months and used the one that gave the best results in the data reported here. The *single segment with groups* model learns $K = 50$ groups in the same way as our proposed model, but is restricted to only have a single time segment, i.e., no time variation in the relative rates of groups. This baseline can be viewed as a probabilistic clustering of the recipients. For the first 3 baselines, only one group exists; the test log-likelihood of these baselines reduce to $LL_{\text{test}} = \sum_{t=1}^{T}\sum_{n=1}^{N_t} \log P(y_{t,n}^{\text{miss}}|\phi)$.

2. 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.

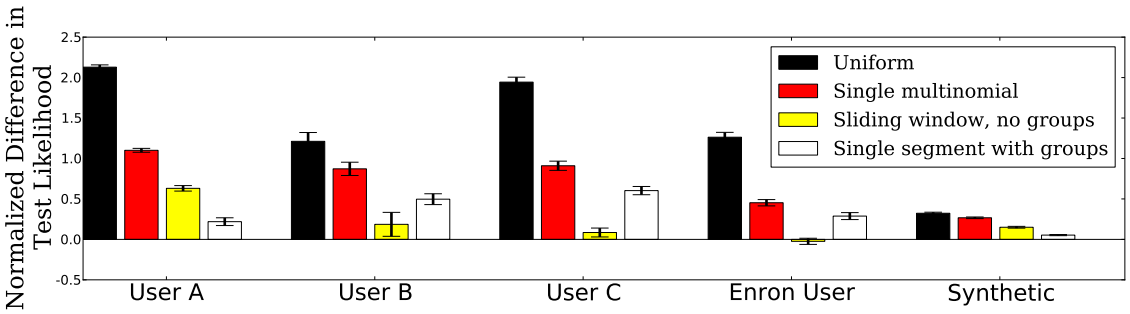


Figure 6: Predictive results from a missing data task for different users.

Figure 6 shows the results for the 4 baselines, relative to our model, across 5 different users. The y-axis is the average test log-likelihood of our model minus the average test log-likelihood of one of the baselines, averaged over 10 different randomly selected missing data sets, where larger positive differences mean that the model outperformed the baseline. Each log-likelihood score was normalized by dividing by the total number of missing recipients for which predictions were made for that user. Users A, B, and C correspond to several years of email data from each of the three authors of this paper, the fourth user is the same Enron user described in Figure 5, and the last user uses the synthetic data set described in Section 5. This plot shows results obtained with 20% of recipients missing at random—almost identical results were obtained with other fractions of missing data (not shown).

The results in Figure 6 for User A and the synthetic user show a clear ordering in terms of the performance of different approaches, with our proposed model being systematically more accurate in predictions than all of the baselines across all 4 data sets. The poor performance of the uniform and multinomial approaches indicate that group information is particularly important, e.g., there can be multiple "active" groups on any given day—this is apparent for the synthetic data in Figures 3 and 4, where we can see significant group overlap both in terms of membership and time. In contrast, Figure 6 shows that the sliding window model is competitive with our proposed model for the other users, indicating that group overlap is not a significant factor in modeling these users. These predictive experiments illustrate that the model can capture useful information from the data, both in terms of temporal variation and group structure, to different degrees for different users.

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 (as in Kleinberg, 2003), superposed on the segments, may be a useful avenue for further exploration. There are also numerous opportunities to extend the modeling of groups. For example, in the present work we fix the number of groups, 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 membership 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-occurrence 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, predictive experiments indicated that the model yields improved predictive accuracy 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.

Acknowledgments

This work was supported in part by an NDSEG Graduate Fellowship (CD, NN), a Google Research Award (PS), and by ONR/MURI under grant number N00014-08-1-1015 (CD, NN, PS).

References

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

M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584. MIT Press, 2002.

D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM Press, 2006.

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

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

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

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

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

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

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

B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*, pages 217–226. Springer, 2004.

R. Malmgren, D. Stouffer, A. Motter, and L. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *PNAS*, 105:18153–18158, 2008.

R. Malmgren, J. Hofman, L. Amaral, and D. Watts. Characterizing individual communication patterns. In *Proceedings of the 15th ACM SIGKDD Conference*, pages 607–616. ACM Press, 2009.

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

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

M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD International Conference*, pages 233–242. ACM Press, 2010.

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

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

C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference*, pages 717–726. ACM Press, 2007.

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

K. Xu, M. Kliger, and A. Hero. Tracking communities in dynamic social networks. In *Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 219–226. Springer-Verlag, 2011.