Topic-Partitioned Multinetwork Embeddings

Peter M. Krafft *
CSAIL
MIT
pkrafft@mit.edu

Juston Moore, Hanna M. Wallach
Dept. of Computer Science
UMass Amherst
jmoore@cs.umass.edu
wallach@cs.umass.edu

Bruce Desmarais
Dept. of Political Science
UMass Amherst
desmarais@polsci.umass.edu

Abstract

We introduce a joint model of network content and context designed for exploratory analysis of email networks via visualization of topic-specific communication patterns. Our model is an admixture model for text and network attributes that uses multinomial distributions over words as admixture components for explaining email text and latent Euclidean positions of actors as admixture components for explaining email recipients. This model allows us to infer topics of communication, a partition of the overall network into topic-specific subnetworks, and two-dimensional visualizations of those subnetworks. We validate the appropriateness of our model by achieving state-of-the-art performance on a prediction task and semantic coherence comparable to that of latent Dirichlet allocation. We demonstrate the capability of our model for descriptive, explanatory, and exploratory analysis by investigating the inferred topic-specific communication patterns of a new email data set, the New Hanover County email corpus.

1 Introduction

The structures of communication networks are critical to collaborative problem solving [1]. Email data sets can help researchers directly observe these communication networks. In addition, given their rich textual detail, existing infrastructure, and widespread usage, email data sets hold the potential to answer many important scientific and applied questions. We introduce a novel Bayesian admixture model specifically intended for analyzing complex email communication networks.

Although there has been some work on modeling email data (e.g. the author-recpient topic model [2]), much of the recent work incorporating both text and network data in probabilistic generative models has focused on networks of documents, such as web pages and the links between them [3] or academic papers and their citations [4]. In contrast to these types of data, email networks are networks of individuals; documents are associated with connections between nodes rather than with nodes themselves. Additionally, much recent work on jointly modeling text and network data has focused on link prediction (e.g., the relational topic model [3]), and does not offer the ability to extract qualitative structural information about the network.

Unlike this previous work we focus on latent space embeddings of email communication networks for the purpose of producing network visualizations that are meaningful, precise, and accessible to practitioners who wish to perform exploratory or descriptive analyses. There are a large number

*Work done at UMass

of existing techniques for network visualization; however, rather than viewing an email network as a single communication network, we treat an email network as a composition of multiple topic-specific networks and attempt to visualize each one individually. Viewing the data in this way is useful because it allows us to examine topic-specific behavior. For example, we can investigate whether there are any breaks in communication about particular topics in order to avoid lack of communication between relevant communities. To date, determining and visualizing topic-specific networks has been almost completely unexplored.

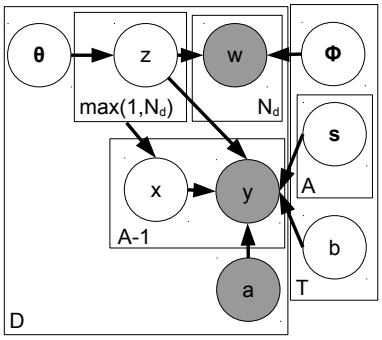
The key challenge in visualizing topic-specific email communication patterns is inferring the topics represented in the emails along with a partition of the entire network into topic-specific subnetworks. Since email data sets are usually unannotated, information about which emails pertain to which topics or which links are best explained by which topic-specific communication patterns is seldom available. In addition, the topics of communication, i.e. groups of words describing each topic, and each topic-specific communication pattern, i.e. each set of tendencies for actors to communicate, are also not typically known themselves a priori.

Our model associates network edges with topics. We view an email data set as a multinetwork consisting of observations of edges between pairs of actors where each edge is an author–recipient pair within an email. According to our model, a word in an email is well-described by a particular topic if that word has high probability in the topic and the topic has high probability in that document. Thus, our model infers groups of words (i.e., topics) that commonly occur together in documents. Furthermore, a group of edges (i.e., a subnetwork) is well-described by a latent space if actors who communicate more frequently in that subnetwork are closer together in that space. Our model identifies each topic's subnetwork by finding edges that are well-described by that topic's latent space and which belong to documents containing that topic.

This model builds upon three previous approaches. Our general strategy is to employ the joint model structure introduced by Correspondence-LDA (Corr-LDA) [5], treating recipients as annotations of emails. As in latent Dirichlet allocation (LDA) [6] and Corr-LDA we use multinomial distributions over words as the admixture components for the text model. We use latent spaces in the sense of Hoff's latent space distance model (LSM) [7], which have not previously been used in the context of mixture or admixture models, as the admixture components of the recipient model. LDA provides us with topics, and LSM provides us with the ability to distinguish topic-specific subnetworks and to embed those networks of actors in two-dimensional space for visualization.

The final contribution of this paper is a new email data set. Due to factors involving personal privacy concerns and private ownership by email service providers, academic researchers rarely have access to email data sets. For example, the Enron corpus [8]—arguably the most-studied email data set—was only released because of a court order. An alternative source of email data is the public record. Such data sets are widely available and can be updated on a continuous basis, yet remain relatively untapped by the academic community. We introduce and analyze a new email data set, relevant to researchers in machine learning as well as to researchers in the social and organizational sciences, consisting of emails between department managers of New Hanover County, North Carolina. The departments comprise the executive arm of government at the county level in North Carolina. In this autonomous local government, the county manager acts as the executive and the individual departments are synonymous with the individual departments and agencies in, for instance, the U.S. Federal government. Specifically, this data set offers a view into the communication patterns of the managers of New Hanover County, but more generally, our analysis serves as a case study in modeling interagency communications in government administration.

With the remainder of this paper we start by providing the mathematical details of our new model and presenting a corresponding inference algorithm. We then demonstrate that the assumptions of a Euclidean network embedding are appropriate for email data. Although our model is intended for exploratory and descriptive analyses, we use link prediction to validate our assumptions. We present state-of-the-art performance on the prediction task. We also verify that our model is able discover topics that are at comparable in quality to those identified by LDA and that is is able to fit various network statistics of our new data set. Finally, we showcase our model's ability to visualize topic-specific communication patterns by inferring topic-specific latent spaces for the New Hanover County email data set. We provide an extensive analysis of these topic-specific latent spaces and demonstrate that they provide accessible visualizations of email-based collaboration, while possessing precise mathematical and probabilistic interpretations.



α	hyperparameter for a symmetric Dirichlet
$\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$	topic proportions for doc. d
β	hyperparameter for a symmetric Dirichlet
$\boldsymbol{\phi}_t \sim \text{Dir}(\beta)$	distribution over word types for topic t
$z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$	topic assignment of token n in doc. d
$w_{dn} \sim \text{Mult}(\boldsymbol{\phi}_{z_{dn}})$	type of token n in doc. d
μ, σ^2	Gaussian hyperparameters
$\boldsymbol{s}_i^{(t)} \sim \mathcal{N}(0, \sigma^2)$	position of actor i in latent space t
$b_t \sim \mathcal{N}(\mu, \sigma^2)$	intercept of latent space t
a_d	author of doc. d
$x_{dj} \sim U(1, \ldots, N_d^*)$	the token assignment of edge j in doc. d
$y_{dj} \sim \text{Bern}(p_{a_d j}^{(z_{dx_{dj}})})$	edge between actors a_d and j in doc. d

Figure 1: The directed graphical model for our new model and a table summarizing the corresponding probabilistic generative process, where $N_d^* = \max(1, N_d)$, and we use σ, the logistic sigmoid function, to map offset Euclidean distances to probabilities. The probability of actor i communicating with actor j in space t is defined as $p_{ij}^{(t)} = \sigma(b_t - \|\boldsymbol{s}_i^{(t)} - \boldsymbol{s}_j^{(t)}\|)$, where $\|\cdot\|$ indicates the ℓ^2-norm.

2 Topic-Partitioned Email Network Embedding

We develop a model for email data that allows us to jointly infer topics present in those emails, a partition of the full communication network into topic-specific subnetworks, and two-dimensional embeddings of those subnetworks. Although this model is more generally applicable to any multinetwork whose edges are annotated with text (e.g. chatroom conversations, discussion threads, or legislative meeting agendas or minutes), we frame our discussion in terms of the email-specific model.

Our model consists of two parts: a text aspect and a network aspect. We model a set of D documents (email messages) $\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\}$ where each w_{dn} "token" is associated with a word "type" that determines the value of w_{dn} in the vocabulary $\mathcal{V} = \{1, \ldots, V\}$. The text aspect of our model is analogous to LDA, a generative probabilistic model for the documents that is an admixture model of "topics", where a "topic" $t = 1, \ldots, T$ is a discrete distribution over the vocabulary described by the probability vector $\boldsymbol{\phi}_t$.

The network aspect, whose admixture components are based on LSM, is slightly more complicated. We assume each document is annotated with two $A \times A$ non-negative integer-valued matrices $\boldsymbol{Y}_p^{(d)}$, which represents document-specific present connections between A actors (authors) $\mathcal{A} = \{1, \ldots, A\}$, and $\boldsymbol{Y}_n^{(d)}$, which represents document-specific absent connections between those actors. We refer to the multinetwork (the non-negative integer-valued matrix) $\boldsymbol{Y}_p = \sum_{d=1}^{D} \boldsymbol{Y}_p^{(d)}$ as the "total present connections". We refer to $\boldsymbol{Y}_n = \sum_{d=1}^{D} \boldsymbol{Y}_n^{(d)}$ as the "total absent connections". Although in the general multinetwork setting, each document's network could have an arbitrary structure, in the case of an email network each email is from a single author $a_d \in \mathcal{A}$, to one or more recipients $R_d \subseteq \mathcal{A} \setminus \{a_d\}$, so in the email-specific case we represent our data as a bipartite graph between emails and recipients, a $D \times A$ matrix $\boldsymbol{Y}$, where $y_{dj} = 1$ if j is a recipient of message d, and $y_{dj} = 0$ otherwise. We assume the author of an email is never a recipient, i.e. $y_{da_d} = 0$ for all d, and omit these edges from the generative procedure.

Our goal is to partition the total present and absent connections $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_n$ into T pairs of multinetworks $\{\boldsymbol{Y}_p'^{(t)}\}_{t=1}^T$ and $\{\boldsymbol{Y}_n'^{(t)}\}_{t=1}^T$, one pair for each topic. We call each such pair a "topic-specific subnetwork" since the edges in each subnetwork contain a subset of the edges from each total network.

The main idea of our model is that each topic-specific subnetwork should be reflective of a topic-specific communication pattern. That is, we view the network aspect of our model as an admixture of "communication patterns", where in our case a "communication pattern" is an $A \times A$ matrix of Bernoulli distribution parameters $\boldsymbol{P}^{(t)}$ that describes the probability each actor will communicate with each other actor. However, rather than using a matrix of free parameters for each communication pattern, we use LSM to constrain these matrices. LSM is a generative model for binary network

data that associates each node in a network with a position in unobserved K-dimensional Euclidean space and explains the relationship between any pair of actors via the pairwise distance between them in the K-dimensional latent space. The probability of two actors connecting according to a particular latent space is a monotonically decreasing function of the distance between those actors in that space—two actors that are closer in the latent space are more likely to share an edge, while two actors that are further apart are less likely.

With A actors, the set of A positions in a latent space is represented by an $A \times K$ matrix $\boldsymbol{S}^{(t)}$. Each $\boldsymbol{P}^{(t)}$ is fully determined by $\boldsymbol{S}^{(t)}$ and an intercept term b_t, so the network aspect of our model can also be viewed as an admixture of latent space models. The primary constraints that LSM enforces on the communication patterns are that they be symmetric and transitive, but it also gives us a natural and internally consistent way to visualize the communication patterns.

We combine these text and network aspects in a full joint probabilistic model. Using a full joint model rather than a multi-stage estimation procedure allows information to flow between the text and network aspects of the model. That is, according to the model recipients of an email contain information about the words present in that email and vice versa. We use the structure introduced by Corr-LDA rather than an exchangeable structure to combine the text and network aspects.

Blei and Jordan developed Corr-LDA as a joint model of images and captions of those images. Corr-LDA introduced an important idea about how to structure joint admixture models of any type of annotated data. Whereas previous joint admixture models of annotated data had used a structure in which all different types of observed data were treated exchangeably (conditionally independent given the admixture proportions of those data), Blei and Jordan suggested that the assignment of annotations to their admixture components should be treated as conditional on the assignments of the main type of data. The problem with the exchangeable model structure is that annotations may be able to be explained by a different set of admixture components than the main data type is explained by, leading to overfitting. The purpose of their modeling choice is to improve discriminative performance, while preserving joint model fit, by guaranteeing that the annotations can only be explained by components also used by the main data type. Although our primary goal is not prediction, in order to trust conclusions drawn using descriptive models (which may not be optimized for predictive performance), such models must also be capable of realistic predictions [9].

Using this type of model structure is especially important for network data since binary edges can easily be explained with just two admixture components, but it is still rarely used in joint models of text and network data. One significant exception is Chang and Blei's relational topic model (RTM) which uses a structure inspired by Corr-LDA for modeling document networks. However, our model is designed for a different purpose than RTM and is not directly comparable to RTM. That is, RTM was designed for networks of documents and is not applicable to email networks since it relies on both the sender and the recipient of each directed edge in the networks it models to be associated with text.

2.1 Generative Process

The graphical model and generative process for our model are shown in Figure 1. Our model assumes that the text of each email is generated via a combination of latent topics and that the recipients of each email are chosen according to a combination of topic-specific latent spaces. Specifically, the "generative story" of our model assumes the following: Each actor has a position drawn from a Gaussian prior in each of T K-dimensional latent spaces, $\{\boldsymbol{S}^{(t)}\}_{t=1}^{T}$. Each latent space has a particular bias towards communication b_t also drawn from a Gaussian prior. Each document has a particular distribution over topics, $\boldsymbol{\theta}_d$. The prior over each element of $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ is assumed to be a Dirichlet and may be either symmetric or asymmetric. The tokens in every document $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ are associated with corresponding topic assignments $\boldsymbol{z}_d = \{z_{dn}\}_{n=1}^{N_d}$, drawn i.i.d. from $\boldsymbol{\theta}_d$, while each token itself is drawn from the topic indicated by its topic assignment $\boldsymbol{\phi}_{z_{di}}$. A Dirichlet prior is placed over each element of $\Phi = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T\}$. We assume these priors are are symmetric as is typical in LDA. Whether or not each of the $A - 1$ possible recipients is present ($y_{dj} = 1$) or absent ($y_{dj} = 0$) in each document is determined by the probability of the author of that document a_d connecting with that recipient j according to the latent space $z_{dx_{dj}}$, where x_{dj} is associated with a random topic that was used by at least one word in the document.

Since certain emails may contain no subject line and no body, when $N_d = 0$ we include a dummy token assignment variable that is not associated with any actual tokens. The edges in such emails are all generated from a single topic, which is essentially as if they had been generated according to a mixture model rather than an admixture model.

The key properties of this model are that $\boldsymbol{z}$ associates topics with latent spaces, $\boldsymbol{x}$ partitions the observed email network into topic-specific subnetworks, and each $\boldsymbol{S}^{(t)}$ is directly visualizable and interpretable. This means that during inference, subnetworks induced by the $\boldsymbol{x}$ partition are simultaneously inferred and embedded in topic-specific Euclidean spaces.

2.2 Inference

To perform inference in our model we integrate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ out of our joint distribution as in collapsed Gibbs sampling for LDA [10], then perform a Metropolis-within-Gibbs MCMC procedure. The sampling algorithm consists of drawing each z_{di} and x_{dj} directly from their full conditional distributions given all of the data and other model parameters, and using Metropolis steps to sample each $\boldsymbol{s}_i^{(t)}$ and b_t. Letting $t = 1, \ldots, T$, the full conditional for z_{di} is given by

$$p(z_{di} = t|\boldsymbol{z}_{-di}, \ldots) \propto \begin{cases} (n_{t|d}^{\backslash di} + \alpha) \frac{n_{w_{di}|t}^{\backslash di} + \beta}{n_{\cdot|t}^{\backslash di} + V\beta} \prod_{j:x_{dj}=i} (p_{a_d j}^{(t)})^{y_{dj}} (1 - p_{a_d j}^{(t)})^{1-y_{dj}}, & \text{if} N_d > 0 \\ \prod_{j \in \mathcal{A} \backslash \{a_d\}} (p_{a_d j}^{(t)})^{y_{dj}} (1 - p_{a_d j}^{(t)})^{1-y_{dj}}, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{z}_{-di}$ contains the token assignments besides z_{di}; the ellipses indicate conditioning on all of the remaining parameters besides those that have been collapsed, the hyperparameters, and the data; and $n_{t|d}^{\backslash di}$, $n_{w_{di}|t}^{\backslash di}$, and $n_{\cdot|t}^{\backslash di}$ are the number of times topic t has been assigned in document d without including z_{di}, the number of tokens of type w_{di} that have been assigned to topic t not including w_{di}, and the total number of tokens assigned to topic t not including w_{di}, respectively. $p_{ij}^{(t)}$ is defined in Figure 1.

Letting $i = 1, \ldots, N_d^*$ the full conditional distribution for x_{dj} is given by

$$p(x_{dj} = i|\boldsymbol{x}_{-dj}, \ldots) \propto (p_{a_d j}^{(z_{di})})^{y_{dj}} (1 - p_{a_d j}^{(z_{di})})^{(1-y_{dj})}.$$

Likewise, both $p(\boldsymbol{s}_i^{(t)}| \ldots)$ and $p(b_t| \ldots)$ are proportional to the full joint distribution. Assuming an improper uniform prior, the full conditional distribution for $\boldsymbol{s}_i^{(t)}$ is

$$p(\boldsymbol{s}_i^{(t)}| \ldots) \propto \prod_{j \in \mathcal{A} \backslash \{i\}} (p_{ij}^{(t)})^{n_{ij1|t}} (1 - p_{ij}^{(t)})^{n_{ij0|t}}$$

where $n_{ij1|t}$ is the number of times a present edge from i to j or from j to i has been assigned to topic t and $n_{ij0|t}$ is the number of times a negative edge has been assigned to topic t. Similarly, the full conditional distribution for b_t is

$$p(b_t| \ldots) \propto \prod_{i,j \in \mathcal{A}: i<j} (p_{ij}^{(t)})^{n_{ij1|t}} (1 - p_{ij}^{(t)})^{n_{ij0|t}}$$

Although we cannot sample from these latter two distributions directly, we can use them for Metropolis-Hastings updates.

3 Data

Another major contribution of this paper is a new email data set: the entire inboxes and outboxes of the county managers of New Hanover County, North Carolina from the month of February, 2011. In this data set there are 30 managers whose inboxes and outboxes are fully observed. There are 1,739 emails between the managers themselves (not including messages from a manager to only him/herself), 8,097 emails authored by the managers, and 30,909 emails in total authored or received by the managers. These 30 managers represent 27 different departments, three of which have special roles: the departments of the County Manager, the Board of Commissioners, and the District Attorney. The other departments oversee some specific parts of the county such as Parks and

Gardens, Youth Empowerment Services, Budget, or Taxes. In all our experiments, we use the fully observed subset of our full email network composed of the emails sent between the managers.

We also validate our model on the Enron email corpus [8], a well-known data set that was collected and publicly released as part of an investigation of the Enron corporation scandal. We use this well-known data set to verify that our model is applicable beyond the data from New Hanover County. Some actors use multiple email addresses in the Enron data set, and the task of entity resolution was beyond the scope of this paper. Therefore, we considered each unique email address to be an individual actor. We chose to look at the most active 50 addresses used in the "from" field of messages, where activity was computed as total degree (number of messages sent or received). We only consider messages from the "_sent_mail", "sent", and "sent_items" folders in order to avoid duplicate messages. In total, we had 8,322 messages between the 50 actors.

We preprocessed both data sets to remove stop words, URLs, quoted text, and signatures that were separated by clear boundaries. We treated the subject of each email as part of its body. We treated the "To" and "Cc" fields equivalently so that a positive edge exists between two actors for each time one is in the "From" field of an email and the other is in the "To" or "Cc" field. We ignored "Bcc" fields and mailing lists.

4 Experiments

Our primary goal in this section is to illustrate the utility of our model as an exploratory and descriptive tool, but we first validate our model using a link prediction task and a topic coherence task. These experiments are designed to show that the model is not overfitting and that the topics learned during joint inference are at least as meaningful as the topics learned by LDA alone. Our model outperforms all our comparison methods on link prediction and achieves topic coherence comparable to LDA and the other baselines. These results suggest that our model can achieve state-of-the-art predictive performance without harming the coherence of the topics it infers, and that our modeling assumptions are reasonable. After discussing these validation experiments, we then focus on the New Hanover County (NHC) data set. We first show using a simulation study that our model can represent network statistics of the NHC corpus. We then conduct an in-depth analysis to showcase the novel capability of our model to visualize topic-specific communication patterns.

4.1 Link Prediction

We use a link prediction task to validate our modeling assumptions and benchmark our model's generalization performance against existing methods. The setup for this task is as follows. For each repetition we obscure the recipients of each email with probability 0.1. When an email is obscured, we treat all $A - 1$ possible edges in that email as missing values, and we then try to recover the true values. Since we do not obscure the text of any emails, this task can be interpreted as predicting the recipients of the obscured emails given the subject and body of those emails. For our model and the comparison methods that involve generative models (everything but the simple network-only baseline), we infer missing edges via Gibbs sampling according to the full conditional distributions of those edges. In all cases, in order to determine the predicted links for a given method on a particular repetition we choose a single random sample from that method's inference procedure. We then compute the F-score for that sample by comparing it to the true values of the missing edges. We average these F-scores across independent repetitions to obtain the final average F-score for each hyperparameter setting of each method.

We initialize all the discrete variables from their priors, we initialize all of the latent positions from K-dimensional standard normal distributions, and we initialize all of the intercepts to the value 10. Using this initial value for the intercepts makes sense for our link function because distance between positions in each latent space can only decrease the probability of actors communicating. For all Metropolis-Hastings proposal distributions we use normal distributions centered around the last sampled value with diagonal proposal covariance of $\max(1, 100/i)$ where i is the iteration number.

We compare to five other methods which we refer to as "*bernoulli*", "*erosheva*", "*lsm*", "*mmsb*", and "*baseline*" (our model is abbreviated as "*tpme*"). *bernoulli* is equivalent to our model except that instead of using latent spaces in its network admixture components, it uses T, $A \times A$ (sym-

metric) Bernoulli probability matrices which directly model the probability of each pair of actors communicating according to each topic-specific network admixture component. The purpose of this comparison is to validate the modeling assumption that two-dimensional Euclidean spaces can accurately summarize communication patterns. *erosheva* was introduced by Erosheva et al. as an alternative formulation of their joint mixed membership model of text and network data [4]. It is similar to *bernoulli*, but it uses the standard exchangeable mixed membership joint model structure rather than that of Corr-LDA, and its admixture components are $A \times 1$ vectors of Bernoulli parameters, one for each possible recipient, rather than $A \times A$ matrices of Bernoulli parameters. This model is more appropriate for link prediction than the primary model of Erosheva et al. since it does not condition on the number of recipients present in an email. The purpose of comparing to *erosheva* is to compare to a related state-of-the-art method for joint modeling of text and network data. *lsm* is simply the latent space distance model by itself. The purpose of comparing to LSM is to show that having a admixture of latent spaces helps us represent our network data. *mmsb* is the mixed membership stochastic blockmodel [11]. *mmsb* uses only network information and uses a model structures that is different from both the standard mixed-membership framework and from that of Corr-LDA. The purpose of comparing to *mmsb* is the same as for comparing to *erosheva*, situating our work within the existing literature. Finally, *baseline* is a network-only baseline in which the probability that $y_{dj} = 1$ is the number of present edges from a_d to j in the training set divided by the number of present edges plus the number of absent edges. The purpose of comparing to *baseline* is to ensure we are doing better than a naïve approach. We use collapsed Gibbs sampling to fit *bernoulli*, *erosheva*, and *mmsb* (using the Gibbs sampling equation proposed by Chang [12] for *mmsb*), and we used Metropolis-Hastings to fit *lsm*.

Based on an earlier optimization of the hyperparameters of LDA alone, *bernoulli*, and *erosheva* we use $\alpha = 1/T$ with the NHC data and $\alpha = 2/T$ with the Enron data. We use $\beta = 0.01$ for all three models and $\gamma = 0.01$ (analogous to β for the network admixture components) for *bernoulli* and *erosheva* on both data sets. These latter hyperparameters matched our LDA optimization and were consistent with those found in previous work [10]. For *mmsb*, we performed a grid search over hyperparameter values.[1] For our model and *lsm* we use improper noninformative uniform priors on the latent positions and intercepts.

Figure 2 shows the results for the NHC and Enron data sets. For our model and all other models that used topics, we show results for all numbers of topics we ran. We used two-dimensional latent spaces for all topic sizes with our model, but we varied the latent space dimension for *lsm*. We plot only the best performing number of admixture components for *mmsb* ($K = 30$) and in terms of predictive F-score. We ran *mmsb* for 5000 iterations and all other MCMC algorithms for 50000 iterations. The results presented were averaged over five repetitions.

Our model outperforms all other baselines. This indicates that we achieve state-of-the-art performance and that our modeling assumptions are reasonable. In particular, since *bernoulli* does not represent transitivity, which is inherent in our full model, the fact that our model outperforms this unconstrained version shows that our assumption of a two-dimensional latent space embedding is reasonable.

4.2 Topic Coherence

Next we show that the topics inferred by our joint model are interpretable. We use average topic coherence [13] as a quantitative measure of the semantic coherence of the topics inferred from our model, LDA alone, *bernoulli*, and *erosheva*. Figure 2 shows that for a variety of topic sizes, the average coherence of topics learned by our model and all of our relevant comparison methods on the NHC and the Enron emails are very similar to those of LDA. These results are encouraging since the purpose of including LDA in our joint model is not to improve the topic model but to automatically learn meaningful labels for our inferred latent spaces, and LDA topic are generally considered semantically meaningful.

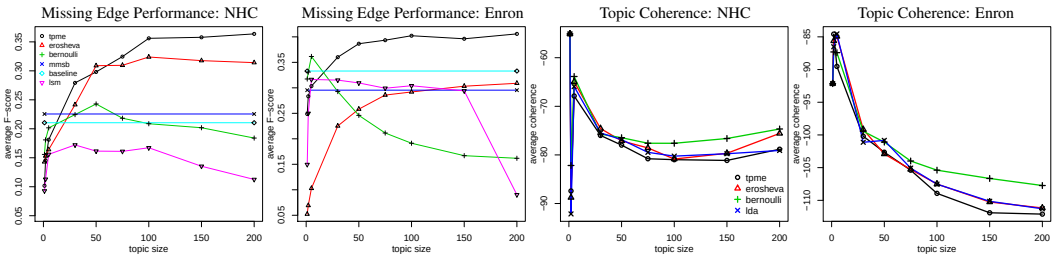


Figure 2: A comparison of average link prediction performance for the NHC data set, the same task for the Enron data set, a comparison of average coherence for the NHC data set, and the same task for the Enron data set. The abscissa values plotted are 1, 2, 5, 30, 50, 75, 100, 150, and 200, which are topic sizes for *tpme*, *erosheva*, *bernoulli*, and *lda*, and the latent dimensions for *lsm*. The legends apply to both like plots.

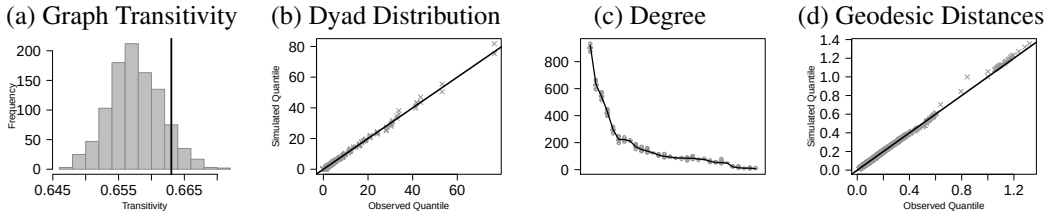


Figure 3: In-sample fit of the models. Panel (a) gives a histogram of the transitivity of the simulated networks, with the transitivity of the observed network located at the vertical black line. A quantile-quantile (QQ) plot comparing the distribution of dyadic intensities in the simulated networks to the observed network is given in panel (b). Panel (c) gives a boxplot of the simulated degree of each manager in the network with managers sorted from highest degree to lowest degree. The black line shows the true vertex degrees. Panel (d) shows a QQ plot comparing the observed and simulated geodesic distance distributions.

4.3 In-Sample Model Fit

In order to explore the NHC data with our model, we select the best performing hyperparameters ($T = 100$) from our link prediction task and train our model on our full data set of emails between managers. To evaluate how well our model fits the data, we simulate 1,000 networks according to the generative procedure of our model given a random sample of our model parameters from their posterior distribution after 40,000 iterations of Gibbs sampling. We then compare the observed values of various network statistics to those simulated according to the model.

Following [14], we look at vertex-wise, dyadic and triadic features of the training network. The dyadic statistic we consider is the geometric mean of the number of emails sent from i to j and the number of emails sent from j to i for each dyad $(i, j) \in \mathcal{A} \times \mathcal{A}$. This is consistent with the suggestion of [15] that dyad intensities in valued networks be measured by the subgraph geometric mean. The vertex-level statistic we consider is the total degree of each manager (in-degree plus out-degree of the total present edge network). The triadic statistic we consider is the generalized graph transitivity, which is a graph-level statistic defined as the sum of all the values of all nonvacuous transitive triples divided by the sum of all values of all the nonvacuous triples, where the value of a triple is the the sum of the two edges participating in that triple [16]. A nonvacuous triple is a triple of nodes (i, j, k) such that $y_{ij} > 0$ and $y_{jk} > 0$ where these values are taken from $\boldsymbol{Y}_p$. A nonvacuous transitive triple is a nonvacuous triple (i, j, k) in which $y_{ik} > 0$. We also compare the observed and simulated geodesic distance distributions, where the distance from actor i to actor j is the reciprocal of the number of emails i sends to j, and the geodesic distance is computed as the shortest path between two managers according to this distance.

The results are depicted in Figure 3. Our model does not perfectly fit the transitivity of the network. However, the observed transitivity is not far enough in the tail of the simulated distribution of

[1]We obtained the best performance using 15 blocks, a symmetric Dirichlet(0.1) prior over group memberships, a Beta(0.1, 0.1) prior over blockmodel diagonal entries, and a Beta(0.01, 0.01) prior over blockmodel off-diagonal entries. We ran the *mmsb* to convergence, which required 1750 iterations.

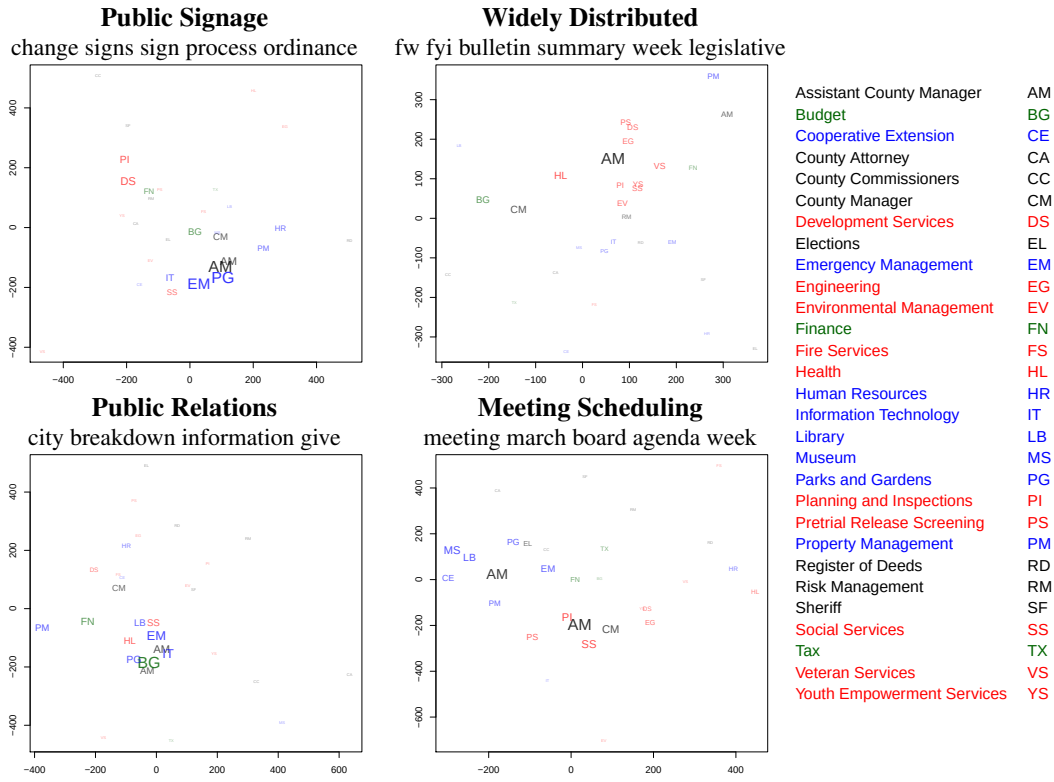


Figure 4: Four example topic-specific latent spaces. The plot for each space is titled with a human-selected label (in bold) for the corresponding topic. Below each label are four or five of the most probable words in the corresponding topic, ordered from most to least probable. The size of each department's acronym indicates how often that department communicates in that space. Acronyms are colored according to their respective "division" in the New Hanover County government organizational chart, which can be found at `www.nhcgov.com/Budget/Documents/FY10-11%20Adopted%20Budget.pdf`. Notice that the acronym "AM" appears twice in all plots because there are two assistant county managers.

networks to warrant additional parameterization of the network, especially since our model fits the other three network statistics well.

4.4 Exploratory Analysis

For our qualitative analysis, we use the same (random) sample of model parameters that we used for our in-sample fit experiment. In Figure 4 we present four topic-spaces that we learn in the NHC data. We select the topics to highlight the insights that can be garnered from topic-space visualizations. Recall that proximity in a topic-specific latent space is directly related to the likelihood that two department managers will include each other in conversations about that topic. Though many other structural characteristics of this communication space may be of interest, we limit our focus to two properties.

First, we examine whether it appears that there are active, disconnected components in the communication space (i.e., modularity), which would indicate that there are different groups discussing the same subjects, but not communicating with each other. Second, through the combined examination of vertex coloration and proximity, we can see whether departments are acting in accordance with the broader organizational structure (i.e., assortativity). Of the four presented, we see evidence of assortativity is in the Widely Distributed and Meeting Scheduling topics. The Public Signage and Meeting Scheduling topics exhibit modularity, with distinct clusters of active nodes. On the other hand, the Public Relations topic, which includes communications with news agencies, is characterized by a single cluster of a wide variety of departments. Lastly, the Meeting Scheduling topic also

displays hierarchical structure with two assistant county managers at the centers of groups consisting of their subdivisions.

Examinations of communication spaces partitioned by communication topic would be very useful for government organizations. First, in regards to assortativity, if organizations find that proximity in communication spaces is generally not related to the official organizational structure, that may serve as a signal that re-organization would be a fruitful effort. If there are particular important topics in which communication is not assortative with respect to the official organization structure, it may be useful to establish inter-divisional structures (e.g., committees). Second, with regard to the search for disconnected components, each component may benefit from efforts to facilitate inter-component communications, possibly drawing on alternative perspectives or foci within a given topic.

5 Conclusions

We have introduced a novel joint admixture model for network data. This model views the communication network between actors as being generated by a combination of topic-specific social processes that can each be summarized by a set of distances between actors. Using this latent space model is additionally useful because it allows us to incorporate information about both "present connections" and "absent connections" within each topic-specific subnetwork. That is, our model allows us to infer topic-specific subnetworks along with interpretable embeddings of these communication patterns. This output is useful in situations when actors display different communication patterns when discussing different topics. In particular, we have shown that this model is useful for substantive exploration of the NHC email data set. Through plots of the topic-specific latent spaces, we are able to (1) analyze the degree to which communication patterns are consistent with the NHC organizational chart and (2) identify groups of government departments who communicate within groups on the same topics, but not across groups.

Our methodology is powerful, intuitive, and generally applicable to any email corpus as well as more generally to multinetworks with text associated with their edges. Any organization that conducts a substantial proportion of its communications via email would find our method useful for summarizing the topics of its internal communications as well as the interactive structure within those topics.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. This work is in supported in part by the National Science Foundation under NSF grant #CNS-0619337. Any opinions, findings conclusions or recommendations expressed here are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] Winter Mason and Duncan J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

[2] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.

[3] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, 2009.

[4] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220, 2004.

[5] D.M. Blei and M.I. Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[8] B. Klimt and Y. Yang. Introducing the enron corpus. In *First conference on email and anti-spam (CEAS)*, 2004.

[9] P.A. Schrodt. Seven deadly sins of contemporary quantitative political analysis. In *106th Annual American Political Science Association (APSA) Meeting and Exhibition, Washington, DC*, 2010.

[10] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.

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

[12] J. Chang. Uncovering, understanding, and predicting links. 2011.

[13] D. Mimno, H.M. Wallach, E.T.M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. EMNLP, 2011.

[14] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 5 2008.

[15] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. Discovering long range properties of social networks with multi-valued time-inhomogeneous models. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[16] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.