Content-Conditioned Hierarchical Latent Space Models for Textual Communication Networks

Matthew Denny [*], **James ben-Aaron** [*], **Hanna Wallach** [*] [†], **and Bruce Desmarais** [*]

[*]University of Massachusetts Amherst, and [†]Microsoft Research NYC

The analysis of complex social structure, and the learning of content from text represent two of the methodological frontiers in computational social science. We present methods that integrate the two. Statistical topic models are widely used to represent content in textual corpora. When documents in a corpus are attributed with senders and recipients, the corpus constitutes text-valued relational or network data (i.e., social structure). We develop a class of models for the topical content and relational structure of communications in which the rate of communication between actors is governed by covariate-conditioned latent space embeddings of the actors, and each embedding is associated with a cluster of topics. We present Markov Chain Monte Carlo methods for inference within the Bayesian framework. This model holds the potential for extensive application to real-world social and organizational relational communication data. To illustrate the model and evaluate its performance, we present an application to internal e-mail communications among managers in 20 North Carolina county governments.

Introduction. Advances in statistical models for textual data have recently been built upon to integrate structural models of networks with models for textual content [5, 4]. In this paper we introduce covariate conditioning and a latent-class topic clustering extension to the topic partitioned multinetwork embeddings (TPME) model introduced by Krafft et al. [4], which situates it as a full generalization of the latent space model (LSM) [3] to text-valued networks. This allows us to make inferences about the content-conditional structure of communication networks, providing an unprecedented model-based window into domain specific sub-structures.

We apply our new generalized content-partitioned multinetwork embeddings (CPME) model to the analysis of e-mail communications among department managers in 20 North Carolina county governments. Utilizing our model extension, we examine the variation in gender mixing patterns across content domains and across counties, buttressing the literature on gender in organizations. This provides a rare look at how gendered network structure varies systematically with the content of professional communication networks. This approach represents a major advance in our understanding of organization communication networks because it allows us to discover varying content-conditional substructures within these networks that a consideration of the aggregate network is likely to miss (an example is provided in Figure 1).

Cluster-Partitioned Multinetwork Embeddings. Here we provide an overview of our update to the TPME model [4] to include latent class topic clustering and edge covariate effects. As in LDA [1], each token is assigned to a "topic" t. Each topic t is also associated with a cluster assignment C_t, where C_t can take one of $C = \{1, ..., c\}$ values. Topics that share a cluster assignment are associated with an $A \times A$ matrix of probabilities $P^{(c)}$ (following the LSM) that a message author a will include recipient r on the message, given that it is about a topic in cluster c.

To associate sender-receiver edges with topic assignments, for each document and each topic, we multiply the proportion of tokens in that document assigned to that topic by empirical edge weights. Thus all edges associated with a particular document are associated with a distribution over topics (describing their content). When we aggregate topics up to their shared latent space assignments, we see variation in the network structure by way of variations in email topical content and receivers. The graphical model for the CPME generative process is shown in figure 2

Inference. For real-world message data , we observe tokens $\mathcal{W}$, authors $\mathcal{A}$, recipients $\mathcal{Y}$ and edge types $\mathcal{I}$, while Φ, Θ, C_t $\mathcal{S} = \{S^{(c)}\}_{c=1}^{C}$, $\mathcal{B} = \{b^{(c)}\}_{c=1}^{C}$, $\Gamma = \{\gamma^{(c)}\}_{c=1}^{C}$, $\mathcal{Z} = \{z^{(d)}\}_{d=1}^{D}$, and $\mathcal{X} = \{X^{(d)}\}_{d=1}^{D}$ are unobserved. Dirichlet-multinomial conjugacy allows us to marginalize out Φ and Θ [1], and sample the remaining unobserved variables from their joint posterior distribution using Markov chain Monte Carlo methods. Metropolis-within-Gibbs can then be used to perform inference. As $z_n^{(d)}$ is a discrete random variable, it can be sampled directly using:

$$P(z_n^{(d)} = t | w_n^{(d)} = \nu, \mathcal{W}_{\setminus d,n}, \mathcal{A}, \mathcal{Y}, C_t, \mathcal{S}, \Gamma, \mathcal{I}, \mathcal{Z}_{\setminus d,n}, \mathcal{X}, \alpha, \beta) \propto$$

$$\begin{cases} \left(N_{\setminus d,n}^{(t|d)} + \frac{\alpha}{T}\right) \frac{N_{\setminus d,n}^{(\nu|t)} + \frac{\beta}{V}}{N_{\setminus d,n}^{(t)} + \beta} \frac{1}{N^{(d)}} \prod_r \left(p^{(c)}_{a^{(d)}r}\right)^{y_r^{(d)}} \left(1 - p^{(c)}_{a^{(d)}r}\right)^{1-y_r^{(d)}} \\ \frac{1}{N^{(d)}} \prod_r \left(p^{(c)}_{a^{(d)}r}\right)^{y_r^{(d)}} \left(1 - p^{(c)}_{a^{(d)}r}\right)^{1-y_r^{(d)}} \qquad \text{for } N^{(d)} = 0 \end{cases} \quad [1]$$

where $\setminus d, n$ indicates a quantity excluding the current token in the current message. New values for the discrete random variable $x_r^{(d)}$, the edge topic assignments may also be sampled directly by comparing their associated edge likelihoods under each cluster latent space.

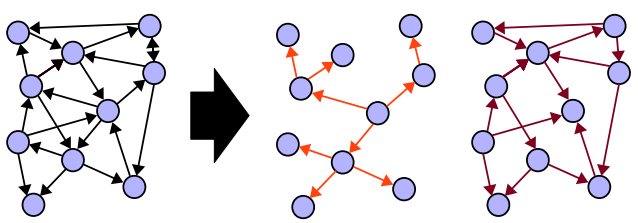

Fig. 1. Different patterns of communication across different domains.

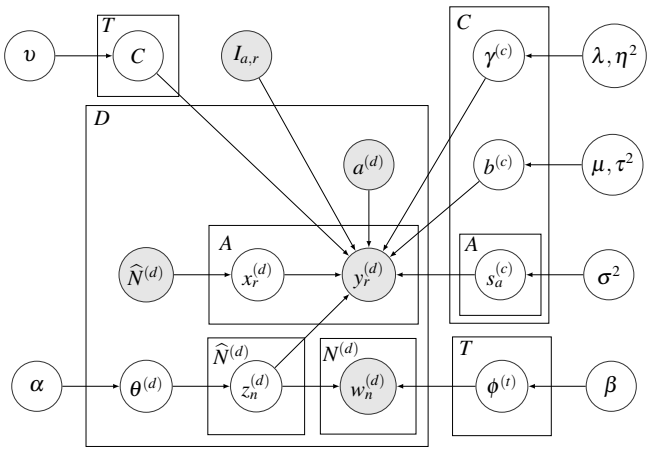

Fig. 2. Graphical model for CMPE generative process.

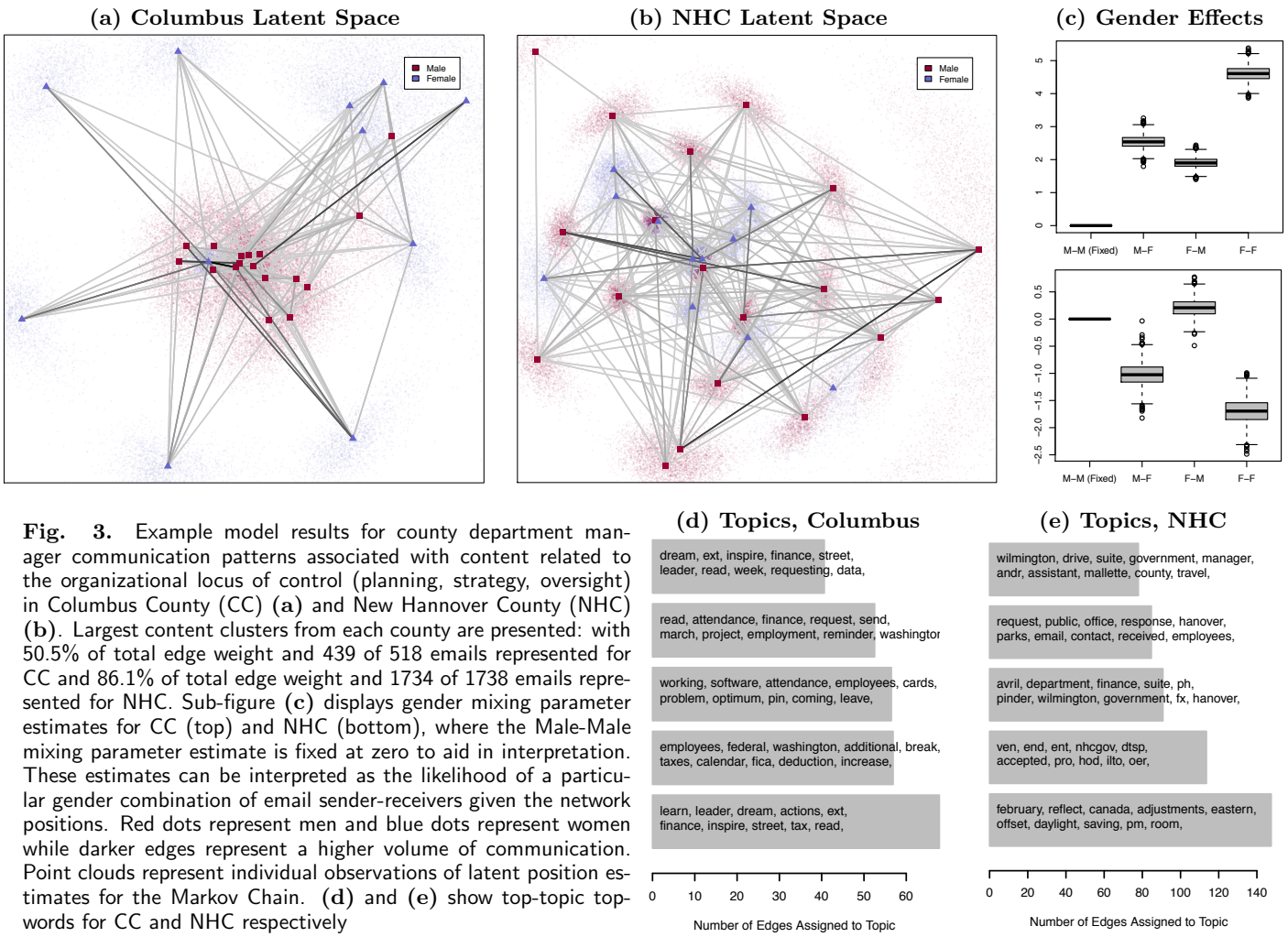

Fig. 3. Example model results for county department manager communication patterns associated with content related to the organizational locus of control (planning, strategy, oversight) in Columbus County (CC) **(a)** and New Hannover County (NHC) **(b)**. Largest content clusters from each county are presented: with 50.5% of total edge weight and 439 of 518 emails represented for CC and 86.1% of total edge weight and 1734 of 1738 emails represented for NHC. Sub-figure **(c)** displays gender mixing parameter estimates for CC (top) and NHC (bottom), where the Male-Male mixing parameter estimate is fixed at zero to aid in interpretation. These estimates can be interpreted as the likelihood of a particular gender combination of email sender-receivers given the network positions. Red dots represent men and blue dots represent women while darker edges represent a higher volume of communication. Point clouds represent individual observations of latent position estimates for the Markov Chain. **(d)** and **(e)** show top-topic top-words for CC and NHC respectively

Topic-cluster assignments may also be updated using Gibbs sampling by constructing a distribution over cluster assignments for each topic by taking the product of edge likelihoods for edges associated with that topic in each cluster latent space. The values for the latent space intercepts $b^{(c)}$, positions $s_a^{(c)}$ and mixing parameter estimates $\gamma^{(c)}$ cannot be sampled directly from their conditional posteriors, but may be obtained using the Metropolis-Hastings algorithm.

Model and Algorithm Validation. Given that the model and inference algorithm for CPME are quite complex, we have taken additional steps to ensure there are no mathematical or coding errors in our model. We have implemented Geweke's "Getting It Right" test (GiR) [2] for posterior simulators which a model will fail if it contains either coding or mathematical errors. Both the LSM and LDA components of our model pass the GiR test with flying colors which gives us confidence in the interpretation of our results.

Results. Figure 3 illustrates preliminary model results for the largest topic clusters in two North Carolina governments. Each model was run for 2,000 iterations with 1,000 iterations of Metropolis Hastings for each iteration of Gibbs sampling. Models were initialized with 50 topics and 10 clusters, but tend to select for 2-4 clusters with any topics assigned to them, indicating a lower dimensional space of communication patterns. To generate estimates of LSM positions, we run the last iteration of the LSM portion of the model until Geweke statistics demonstrate convergence on greater than 90% of all parameters (5M iterations). Our results indicate that men tend to dominate the locus of control within these organizations, but that this pattern is uneven across counties, and dependent on organizational positions occupied by women.

ACKNOWLEDGMENTS. This work was supported by US National Science Foundation Grant CISE-1320219 (Hanna Wallach and Bruce A. Desmarais, PIs)

References

1. DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
2. John Geweke. Getting It Right. *Journal of the American Statistical Association*, 99(467):799–804, September 2004.
3. Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, December 2002.
4. Peter Krafft, Juston Moore, Bruce A Desmarais, and Hanna Wallach. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems Twenty-Five*, 2012.
5. Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272, October 2007.