Reading between the Emails: Gendered Patterns of Communication in Local Government

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

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

We conduct an analysis of the relationship between gender and communication patterns in a sample of 17 North Carolina county governments. We apply an extension of the recently developed topic-partitioned multinetwork embeddings model to infer the content-conditional structure of manager-to-manager email communication in these county governments. We provide results which illustrate that while aggregate patterns of communication among department managers do not display significant gender bias, the content of communication and positions held by men and women differ significantly. We also find that some previously studied institutional-level factors seem to matter for the gendering of communication, and that women seem to be excluded from the locus of control in these organizations.

Introduction

Researchers have recently begun to integrate structural models of networks with models for textual content [4, 3]. In this paper we introduce two extensions to the topic partitioned multinetwork embeddings (TPME) model introduced by Krafft et al. [3], and apply this model to the analysis of a large cross-organizational email corpus. Specifically, we extend the work of Krafft et al. to include actor-level covariates such as gender or organizational position, and latent-class topic clustering. These extensions situate our new model as a full generalization of the latent space network model [2] to text-valued networks. Our approach allows us to discover varying content-conditional sub-structures within the email networks we study, that we are likely to miss when only examining the aggregate network structure (an example is provided in Figure 1).

We use our new model to understand how gender is related to the patterns of communication between department managers in 17 North Carolina county governments, and how these patterns vary with message content. This new dataset comprises almost 18,000 emails sent between 362 county department managers (e.g. health, budget, HR) during overlapping 3-month periods in 2013. Importantly, we were also able to gather covariate data for each manager, such as their gender and formal position in the organization, allowing us to make inferences about the relationship between these covariates and the observed network structure. Utilizing our model, we examine the variation in the propensity for men and women to send, and be included as recipients of emails, across content domains and counties. We find that while aggregate patterns of communication among department managers do not display significant gender bias, the content of communication and positions held by men and women differ significantly.

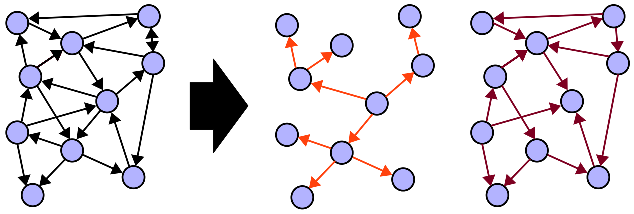


Fig. 1. Different patterns of communication across different domains.

A Model for Communication Networks

Here we provide a brief overview of our extension to the TPME model [3]. We focus on the ways in which our new model allows us to make inferences about the gender-specific patterns of communication in organizations, and how these patterns vary with the content of communication. Our model follows [3] by integrating the latent space network model [2] with a statistical topic model (latent Dirichlet allocation [1]) to jointly model the content and structure of a communication network.

At a high level, our model assumes the following generative process for a message sent across the communication network. First, we sample the content of the message following the generative process for latent Dirichlet allocation. Then, for a given message sender, we sample whether each other actor in the network is a receiver of that message following the latent space network model. However, different topics (or clusters of topics in our extension) are associated with different latent spaces, so how likely each actor is to be a receiver of a particular message is dependent on the topical content of that message. We review this process in greater detail below.

The message content is assumed to be generated via latent Dirichlet allocation. Under this model, we assume that all unique words in our vocabulary are associated to varying degrees with each of T latent "topics". Each topic is then a distribution over unique words (or "word types"), $\boldsymbol{\phi}^{(t)}$. For example "doctor", "virus", and "medicine" might all have high probability in a topic about hospitals, while "cat" and "music" might have low probability in that topic. We further assume that each topic distribution is drawn from a Dirichlet prior:

$$\boldsymbol{\phi}^{(t)} \sim \text{Dirichlet}(\beta, \boldsymbol{n}) \text{ for } t = \{1, ..., T\} \qquad \textbf{[1]}$$

The concentration parameter β controls how peaky this distribution is (how few word types will have high probability in a given topic), and the base measure $\boldsymbol{n}$ controls the bias in the degree to which word types will be assigned high probability in topics generally. For example, "patient" might generally have a higher probability than most word types in a corpus of documents about healthcare.

To generate each of the D documents (messages) in our corpus, we first draw a document-specific distribution over topics

$$\boldsymbol{\theta}^{(d)} \sim \text{Dirichlet}(\alpha, \boldsymbol{m}) \text{ for } d = \{1, ..., D\} \qquad \textbf{[2]}$$

Here, the concentration parameter α controls how topically specific the documents generated are, and $\boldsymbol{m}$ controls the bias in how likely certain topics are to appear. Once we have sampled our document-specific distribution over topics we can then generate each of N tokens (words) in the document via a two step process. First we draw the latent topic assignment for that token from the document specific distribution over topics.

$$z_n^{(d)} \sim \boldsymbol{\theta}^{(d)} \text{ for } n = \{1, ..., N\} \qquad \textbf{[3]}$$

Then we draw the word type of that token from the topic specific distribution over word types.

$$w_n^{(d)} \sim \boldsymbol{\phi}^{(z_n^{(d)})} \text{ for } n = \{1, ..., N\} \qquad \textbf{[4]}$$

We proceed in this manner for all tokens in all documents.

In our extension of TPME, we further assume each topic is uniquely associated with one of C clusters, sampled from a discrete uniform distribution.

$$C_t \sim \text{Discrete Uniform}(1, C) \text{ for } t = \{1, ..., T\} \qquad [\mathbf{5}]$$

The intuition is that different broad content areas of communication (e.g. "planning", "sports","meeting planning") – each of which is a collection of more specific topics – will imply a different pattern of communication. Importantly, under this model, a message can be about a number of topics, and these topics can be associated with different clusters. Each cluster is associated with a different pattern of communication, so the receivers for any particular message must be sampled following an add-mixture of several underlying communication patterns. For example, if a department manager in an organization sent an email message to schedule a budget meeting, they would likely include both staff whose job includes setting up meetings, and staff who needed to provide input on the budget as recipients.

Finally, for a given cluster c, the probability that an actor is selected as a recipient of a particular message is specified by the latent space network model [2]. Under this model, we assume there is some baseline propensity to include message recipients on any email, which is governed by an intercept parameter.

$$b^c \sim \text{Normal}(\mu, \tau^2) \qquad [\mathbf{6}]$$

For example, messages about sensitive HR matters will probably include fewer recipients than messages announcing a department party. Second, some attributes of the sender and potential receiver (which we assume are observed) may also affect the probability of that actor being a recipient. The effect that each of these L different attributes have on the probability of an actor being included as a message recipient can then be parameterized by a vector:

$$\gamma_l^c \sim \text{M. V. Normal}(\lambda, \eta^2) \text{ for } l = \{1, ..., L\} \qquad [\mathbf{7}]$$

For example, there is a great deal of social science literature showing that people tend to preferentially communicate with others of the same gender. Our model could generate this propensity by sampling positive parameters for Male-Male and Female-Female communication, and negative parameters for Female-Male and Male-Female communication.

The additional variation (not captured by covariate effects) in the probability that a message is sent between two actors is governed by how close they are in a k dimensional "latent social space". This latent space captures all un-modeled factors associated with the propensity for two actors to form a tie. In a typical social network this might include difficult or impossible to measure quantities like how nice a person is, or whether two people have "chemistry". Additionally, it may capture observable traits that the researcher was unable to collect data on (e.g. sexual orientation, or ethnicity). This makes the latent positions difficult to interpret when including covariates in the model, so great care should be taken when doing so. To capture these latent positions, each actor a is assumed to have some position in the k-dimensional latent space

$$\mathbf{s}_a^c \sim \text{M. V. Normal}(0, \sigma^2) \text{ for } a = \{1, ..., A\} \qquad [\mathbf{8}]$$

and the probability that they send a message to an actor $r \neq a$ is decreasing in their latent distance from r. Thus the probability of actor r being a recipient of a message from actor a under this model is:

$$P(y_{a,r}^c = 1) = \text{logit}^{-1}\left(b^c + \sum_L \left[\mathbf{X}_{a,r}^l \gamma_l^c\right] - |\mathbf{s}_a^c - \mathbf{s}_r^c|\right) \qquad [\mathbf{9}]$$

where $\mathbf{X}^l$ is a matrix recording the type of an edge for attribute l that would be sent between a and r (e.g. Male to Female or employee to supervisor). Edge values for an individ-

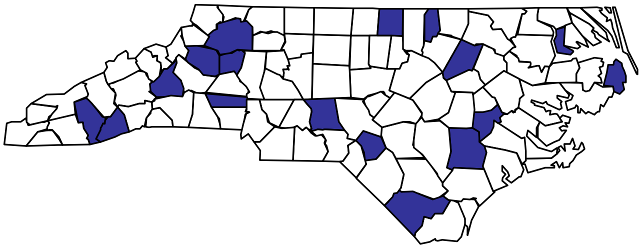


Fig. 2. North Carolina county map.

Table 1. Participating county email statistics. **Mgrs.** is the total number of department managers in a county, **Female** is the number of female managers in that county, **Internal** is the number of emails sent between managers in a county, and **Total** is the total number of emails sent and received by department managers in each county in our sample. Note that in this study, we only make use of the internal email data. Some email **Total**'s are omitted due to challenges in determining which emails (not sent by managers) were valid in these counties.

County	Mgrs.	Female	Internal	Total
Alexander	21	9	907	11,924
Caldwell	20	8	121	
Chowan	23	11	2,027	11,737
Columbus	24	10	920	12,707
Dare	27	12	2,247	
Duplin	27	14	1,914	
Hoke	24	11	1,106	5,565
Jackson	24	6	1,499	
Lenoir	20	5	560	10,499
Lincoln	22	7	573	8,727
McDowell	17	5	326	3,494
Montgomery	18	10	680	2,465
Nash	19	8	1,147	9,133
Person	21	9	1,491	14023
Transylvania	20	4	1,857	14,088
Vance	18	8	185	4,349
Wilkes	17	2	303	8,443
Totals:	362	139	17,863	117,154

Table 2. Email statistics by gender.

	Male	Female
Proportion of All Managers in Sample	61.6%	38.4%
Total Emails Sent	10,771	7,092
Average Emails Received Per Manager	70.8	71.6
Average Number of Recipients Per Email Sent	1.45	1.43
Proportion of Emails Recipients of Same Gender	60.7%	37.8%
Proportion of Emails Sent to County Manager	24.4%	17.9%

[1] The inference algorithm is currently implemented in a beta version as an R package, and is available here: github.com/matthewjdenny/ContentStructure

[2] A technical report detailing the email cleaning and preprocessing procedure is available upon request.

[3] These counties are statistically indistinguishable from the general population of counties in North Carolina on a number of demographic dimensions.

ual message are then sampled via Bernoulli trials. However, the probability that $y_{a,r} = 1$ for a given message d may be dependent on multiple latent spaces (because documents can be about multiple topics). The weight given to $P(y_{a,r}^{(c)})$ for each cluster is determined by the proportion of tokens in the message that are assigned to topics associated with cluster c.

$$P(y_{a,r}^{(d)} = 1) = \sum_{C} P(y_{a,r}^{c} = 1) \times * \qquad \textbf{[10]}$$

where $*$ is the proportion of tokens in document d assigned to topics associated with cluster c.

Given that we do not directly observe the generative process, the object of inference becomes the posterior distribution of our model parameters given the data. This problem is analytically intractable, so we must approximate the posterior distribution via Markov chain Monte Carlo methods. We perform inference for this model via block Metropolis within Gibbs sampling[1]. A further discussion of this model will be provided in a related paper, but this model provides a powerful and flexible framework to investigate the content-conditional gendered patterns of communication in an organization.

Data

To better understand how gender is related to patterns of communication within government bureaucratic organizations, we examine overlapping three-month samples of email messages sent between department managers in 17 North Carolina county governments. The data used in this study were collected in 2013 via a series of FOIA requests, as part of a transparency-by-conformity field experiment involving all 100 North Carolina county governments. We received approximately 500,000 total emails from the 17 participating county governments, of which approximately 120,000 were classified as not spam after an intensive data cleaning and validation process[2]. Figure 2 highlights those counties which did respond to our requests for data[3].

We also collected detailed metadata on the position and gender of the 362 department managers in our sample. We display some basic descriptive statistics for each county in Table 1. Table 2 reports some additional manager-to-manager email communication statistics by gender. These statistics are aggregated across all departments in all counties. Note that emails may have multiple recipients and thus we distinguish between the number of unique emails sent and the number or proportion of recipients.

Analysis

To begin understanding the relationship between gender and the patterns of communication within these county governments, we can look at the descriptive statistics in Table 2. We see that women make up 38% of managers in our sample and they send and receive 39-40% of emails. Importantly, in aggregate, neither men nor women show a strong preference for within gender communication. Therefore, if a preference for same-gender communication exists within these organizations, aggregate level statistics do not capture it. However, these basic descriptive statistics ignore both the formal positions of email senders and recipients, and the email content.

To delve deeper into these patterns of communication, we present four heat maps in Figure 3. These heat maps record the number of emails sent to, and received by department managers given that they were male (female). As we can see, there are significant differences in the number of emails received by male and female managers working in the same department. Most striking is the almost complete absence of emails sent by the county manager to female department managers, whereas male department managers receive the largest proportion of their incoming emails from the county manager. For example, while male finance managers received approximately 40% of their emails (485) from the county manager across all counties in our sample, female finance managers only received 6 emails from the county manager across our entire sample. This differential could reflect gender homophily in communication, given that 15 out of the 17 county managers in our sample are male. However, it may also be linked to the consistent finding in the gender-in-organizations literature that women are excluded from the "locus of control" in organizations.

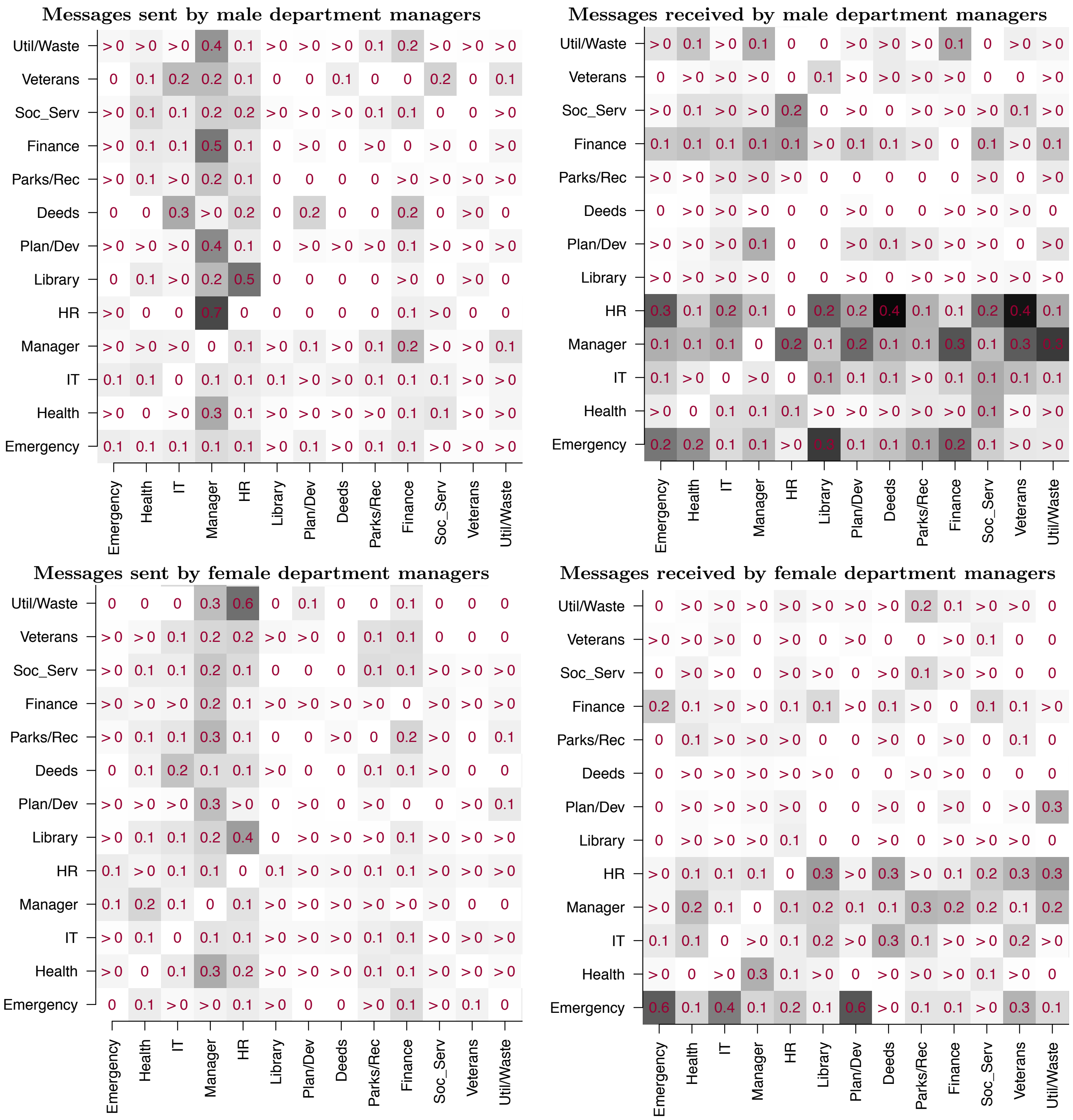
Model-Based Inference. Our initial analysis of the aggregate patterns of interaction in these organizations indicates that there are likely differences in the way department managers in the same organizational position communicate, by gender. It is therefore likely that different content areas will also exhibit different patterns of communication by gender, making this a good application for our model. In the preliminary analysis presented below, we relate the gender-mixing parameters estimated by our model for each county, to the number of department managers in that county (a proxy for organization size). Before we present the results of this analysis, we first briefly discuss our model specification and provide an example of the output it produces.

Our model requires that a number of hyper-parameters be selected by the researcher, prior to performing inference. In particular, we must select the number topics, number of clusters, and topic model hyper-parameters[4], which we hold constant across counties. In addition, we fixed the male-male mixing parameter at zero for our analysis, to aid in directly interpreting the other gender mixing parameters. We selected 40 topics and 4 clusters, to provide reasonable granularity in capturing the content of communication, while improving the interpretability of the latent space model results by constraining the number of possible patterns of communication. This choice has a practical advantage of ensuring that enough data will be available to fit each cluster's latent space with reasonably low uncertainty in the parameter estimates.

The model also requires us to select the number of iterations for our MCMC sampler. In the first step of inference, we alternate between one iteration of Gibbs sampling for the topic model parameters, and 1,000 iterations of Metropolis Hastings sampling for the latent space parameters, as the Metropolis Hastings algorithm explores the parameter space much more slowly than Gibbs sampling. We did this for a total of 4,000 iterations of Gibbs sampling, until Geweke statistics indicated convergence in the un-normalized topic model log likelihood for all counties. We then ran the latent space component of our model for an additional 10,000,000 iterations, holding the topic model parameters fixed, to ensure that all latent space parameter estimates had converged.

Figure 5 depicts example output from one topic cluster estimated using data from Dare county, North Carolina. Our data collection window happened to overlap with Hurricane Sandy, and this cluster of topics clearly reflects storm related email communication. The mixing parameter estimates for this topic cluster indicate that women tend to send emails about these topics to both men and other women at a significantly higher rate than men send emails to men, given their latent positions. The residual content-conditional social struc-

[4] We used uniform base measures $\mathbf{m}$ and $\mathbf{n}$, and set $\alpha = 1$ and β equal to 0.01 times the length of the vocabulary for each county. This is standard practice in the literature using latent Dirichlet allocation and provides good performance.



Messages sent by male department managers

	Emergency	Health	IT	Manager	HR	Library	Plan/Dev	Deeds	Parks/Rec	Finance	Soc_Serv	Veterans	Util/Waste
Util/Waste	> 0	> 0	> 0	0.4	0.1	> 0	> 0	> 0	0.1	0.2	> 0	> 0	> 0
Veterans	0	0.1	0.2	0.2	0.1	0	0	0.1	0	0	0.2	0	0.1
Soc_Serv	> 0	0.1	0.1	0.2	0.2	> 0	> 0	> 0	0.1	0.1	0	0	> 0
Finance	> 0	0.1	0.1	0.5	0.1	0	> 0	0	> 0	0	> 0	0	> 0
Parks/Rec	> 0	0.1	> 0	0.2	0.1	0	0	0	0	> 0	> 0	> 0	> 0
Deeds	0	0	0.3	> 0	0.2	0	0.2	0	0	0.2	0	> 0	0
Plan/Dev	> 0	> 0	> 0	0.4	0.1	0	> 0	> 0	> 0	0.1	> 0	> 0	> 0
Library	0	0.1	> 0	0.2	0.5	0	0	0	0	> 0	0	> 0	0
HR	> 0	0	0	0.7	0	0	0	0	0	0.1	> 0	0	0
Manager	> 0	> 0	> 0	0	0.1	> 0	0.1	> 0	0.1	0.2	> 0	> 0	0.1
IT	0.1	0.1	0	0.1	0.1	0.1	> 0	> 0	0.1	0.1	0.1	> 0	> 0
Health	> 0	0	> 0	0.3	0.1	> 0	> 0	> 0	> 0	0.1	0.1	> 0	> 0
Emergency	0.1	0.1	0.1	0.1	0.1	> 0	0.1	> 0	0.1	0.1	> 0	> 0	> 0

Messages received by male department managers

	Emergency	Health	IT	Manager	HR	Library	Plan/Dev	Deeds	Parks/Rec	Finance	Soc_Serv	Veterans	Util/Waste
Util/Waste	> 0	0.1	> 0	0.1	0	0	> 0	> 0	> 0	0.1	0	> 0	> 0
Veterans	0	> 0	> 0	> 0	0	0.1	> 0	> 0	> 0	> 0	0	0	> 0
Soc_Serv	> 0	0.1	> 0	> 0	0.2	0	> 0	0	> 0	> 0	> 0	0.1	> 0
Finance	0.1	0.1	0.1	0.1	0.1	> 0	0.1	0.1	> 0	0	0.1	> 0	0.1
Parks/Rec	> 0	> 0	> 0	> 0	> 0	0	0	0	0	0	> 0	0	> 0
Deeds	0	> 0	> 0	> 0	0	0	> 0	0	0	> 0	> 0	> 0	0
Plan/Dev	> 0	> 0	> 0	0.1	0	0	> 0	0.1	> 0	> 0	> 0	0	> 0
Library	> 0	> 0	> 0	> 0	0	0	> 0	0	> 0	> 0	> 0	> 0	> 0
HR	0.3	0.1	0.2	0.1	0	0.2	0.2	0.4	0.1	0.1	0.2	0.4	0.1
Manager	0.1	0.1	0.1	0	0.2	0.1	0.2	0.1	0.1	0.3	0.1	0.3	0.3
IT	0.1	> 0	0	> 0	0	0.1	0.1	0.1	> 0	0.1	0.1	0.1	0.1
Health	> 0	0	0.1	0.1	0.1	> 0	> 0	> 0	> 0	> 0	0.1	> 0	> 0
Emergency	0.2	0.2	0.1	0.1	> 0	0.3	0.1	0.1	0.1	0.2	0.1	> 0	> 0

Messages sent by female department managers

	Emergency	Health	IT	Manager	HR	Library	Plan/Dev	Deeds	Parks/Rec	Finance	Soc_Serv	Veterans	Util/Waste
Util/Waste	0	0	0	0.3	0.6	0	0.1	0	0	0.1	0	0	0
Veterans	> 0	> 0	0.1	0.2	0.2	> 0	> 0	0	0.1	0.1	0	0	0
Soc_Serv	> 0	0.1	0.1	0.2	0.1	0	0	0	0.1	0.1	> 0	> 0	> 0
Finance	> 0	> 0	> 0	0.2	0.1	> 0	> 0	> 0	> 0	0	> 0	> 0	> 0
Parks/Rec	> 0	0.1	0.1	0.3	0.1	0	0	> 0	0	0.2	> 0	0	0.1
Deeds	0	0.1	0.2	0.1	0.1	> 0	0	0	0.1	0.1	> 0	0	0
Plan/Dev	> 0	> 0	> 0	0.3	> 0	0	> 0	0	> 0	0	0	> 0	0.1
Library	> 0	0.1	0.1	0.2	0.4	0	> 0	> 0	> 0	0.1	> 0	> 0	> 0
HR	0.1	> 0	0.1	0.1	0	0.1	> 0	> 0	0.1	0.1	> 0	> 0	> 0
Manager	0.1	0.2	0.1	0	0.1	> 0	> 0	> 0	> 0	> 0	> 0	0	0
IT	> 0	0.1	0	0.1	0.1	> 0	> 0	> 0	0.1	0.1	> 0	> 0	> 0
Health	> 0	0	0.1	0.3	0.2	> 0	> 0	> 0	0.1	0.1	> 0	> 0	> 0
Emergency	0	0.1	> 0	> 0	0.1	> 0	0	0	> 0	0.1	> 0	0.1	0

Messages received by female department managers

	Emergency	Health	IT	Manager	HR	Library	Plan/Dev	Deeds	Parks/Rec	Finance	Soc_Serv	Veterans	Util/Waste
Util/Waste	0	> 0	> 0	> 0	> 0	> 0	> 0	> 0	0.2	0.1	> 0	> 0	0
Veterans	> 0	> 0	> 0	0	> 0	0	> 0	0	0	> 0	0.1	0	0
Soc_Serv	0	> 0	> 0	0	> 0	> 0	0	> 0	0.1	> 0	> 0	0	0
Finance	0.2	0.1	> 0	> 0	0.1	0.1	> 0	0.1	> 0	0	0.1	0.1	> 0
Parks/Rec	0	0.1	> 0	> 0	> 0	0	0	> 0	0	> 0	0	0.1	0
Deeds	0	> 0	> 0	> 0	> 0	> 0	0	0	> 0	> 0	0	0	0
Plan/Dev	0	> 0	> 0	> 0	> 0	0	0	> 0	0	> 0	0	> 0	0.3
Library	0	> 0	> 0	> 0	0.1	0	0	> 0	0	> 0	> 0	> 0	0
HR	> 0	0.1	0.1	0.1	0	0.3	> 0	0.3	> 0	0.1	0.2	0.3	0.3
Manager	> 0	0.2	0.1	0	0.1	0.2	0.1	0.1	0.3	0.2	0.2	0.1	0.2
IT	0.1	0.1	0	> 0	0.1	0.2	> 0	0.3	0.1	> 0	> 0	0.2	> 0
Health	> 0	0	> 0	0.3	0.1	> 0	0	> 0	> 0	> 0	0.1	> 0	0
Emergency	0.6	0.1	0.4	0.1	0.2	0.1	0.6	> 0	0.1	0.1	> 0	0.3	0.1

Fig. 3. Heat maps depicting the number of emails sent from the row department to the column department aggregated across counties and by gender. Departments were also hand coded into one of 25 different categories based on given titles to group departments that perform a similar function. The smallest 11 departments by email volume are omitted from these tables for readability but they display a similar pattern of communication.

ture captured by the latent positions indicates that we have likely omitted some salient covariates (whether a person was the emergency manager or the county manager) as both of these managers sent and received a large number of emails in preparation and response to the storm, and are thus placed close to the center of the latent space.

Having illustrated the essential functionality of our model, we now examine the relationship between gender homophily and organization size across our sample of communication networks. In this preliminary analysis, we aggregate together the mixing parameters estimated for all clusters in all counties. We then examine how each individual mixing parameter (male-female, for example) varies in sign and magnitude with the number of department managers in the county it is asso-

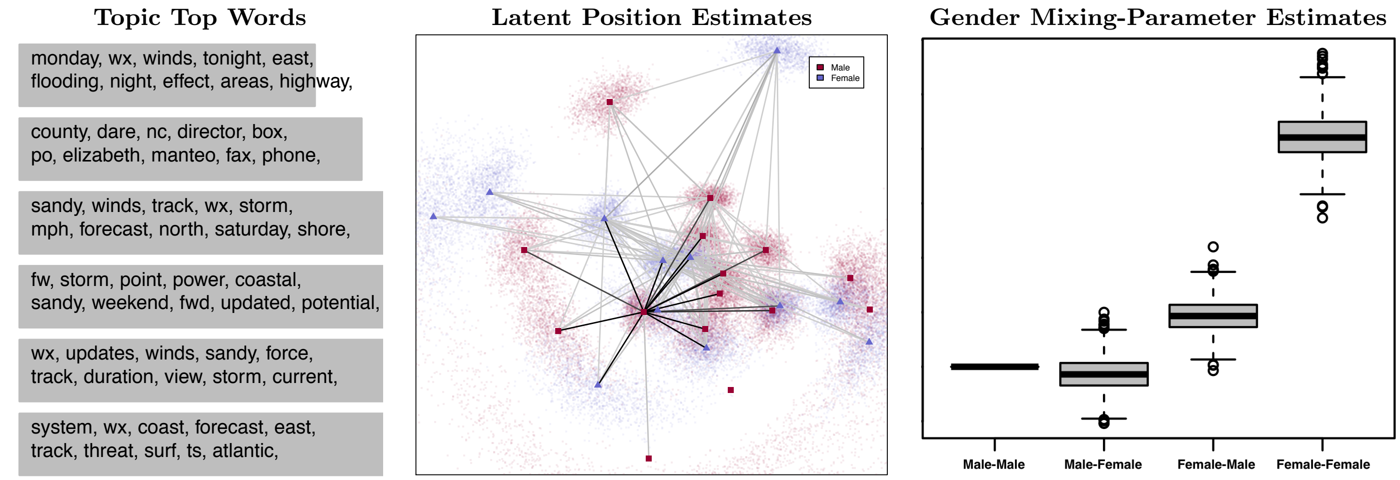


Fig. 5. Model output for selected topic cluster from Dare county. The left-most figure represents the 10 most likely words in the 5 most likely topics (with the most likely topic at the bottom). The middle plot provides latent position estimates for all managers in this county, with female managers represented by blue triangles, and male managers represented by red squares. Point clouds represent 100% credible intervals for latent position estimates, and the darkness of the edge represents its weight. The box plot on the right depicts mixing parameter estimates.

ciated with. This approach allows that what might be driving differences in mixing parameter estimates between clusters and counties is simply differences in what is being communicated, reducing some of the potential for content to confound the effect of department size on gender mixing.

Figure 4 plots the number of managers in a county government (x-axis) against a particular mixing parameter estimate (male-female and female-male) for each of the 4 clusters associated with each of the 17 counties in our sample. Each parameter estimate (dot) is sized by the amount of edge weight assigned to it, to highlight the contribution of more frequent communication patterns. Plots also include LOWESS curves with 95% confidence bounds. We find that heterophilous communication (male-female and female-male) among department managers tends to be most pronounced in mid-sized county governments in our sample. One potential explanation for this finding is that in small organizations, there may not be a great deal of cross-gender communication because the low number of female managers (in absolute terms) encourages group-cohesion among the minority female managers. Additionally, when the organization is large, the increased access (in absolute terms) to communication partners of the same gender could support a gender-homophily explanation for both male and female managers communication patterns.

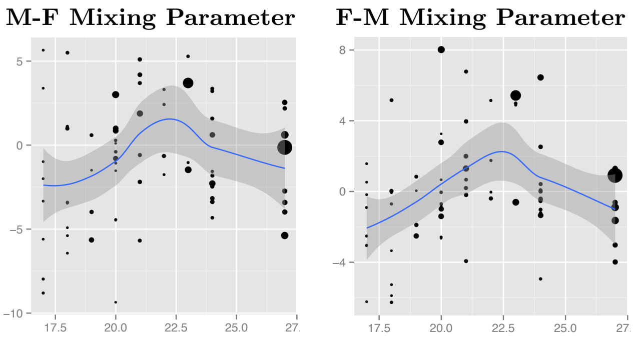


Fig. 4. Gender Mixing parameters plotted against the number of department managers, for each of four topic clusters associated with each county.

Discussion

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. 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.
3. Peter Krafft, Juston Moore, Bruce A Desmarais, and Hanna Wallach. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems Twenty-Five*, 2012.
4. 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.