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

In this paper, we study the role of gender in local government organizations. We analyze email data from seventeen county governments in North Carolina to understand the relationship between a department manager's gender and the emails they send and receive. First, we use descriptive statistics to identify aggregate gender-homophilous and gender-heterophilous patterns from the numbers of emails sent and received by all department managers. In contrast to previous research, we find no strong evidence of either gender homophily or heterophily. To investigate this finding, we therefore analyze department-to-department communication patterns. Here, we do find evidence of gender bias, with some departments exhibiting gender homophily and others exhibiting heterophily. Finally, to determine the extent to which these patterns are driven by communication content, we use a recently developed latent variable model to identify topic-specific communication subnetworks for each county. We find differing degrees of gender homophily and heterophily for different topics of communication. From a policy perspective, these findings suggest that a gender-equitable working environment cannot be created by hiring decisions alone, as gender bias in communication still exists independent of the positions held by men and women.

Gender in Organizations

Researchers have observed and documented gender bias in both the public and private sectors. Women often receive lower pay, hold less prestigious positions, have reduced opportunities for advancement, and are excluded from decision-making coalitions [5, 3, 12, 2, 8]. As a result, organizations that aspire to a just, efficient, and sustainable culture strive to provide men and women with equal treatment in the workplace [10]. In practice, however, these organizations and the researchers who study them have found it hard to fully understand the day-to-day causes and extent of gender bias. The limited availability of primary-source data means that most research is based on small-scale observational, ethnographic, or self-reported data [e.g., 6, 1, 9]. Since these data sources are often restricted in scope and can be biased by subjective assessments, their use in understanding gender bias is limited.

In this paper, we take a different approach and instead base our analyses on email communication data. We focus specifically on local government organizations and seek to understand the role of gender in communication at the department manager level. With the increasing use of electronic communication in the workplace, and the rise of transparency initiatives within government, scholars can now use public records requests to gather primary-source data about government organizations. Moreover, for many government organizations, such requests even extend to emails. We draw upon this resource to construct an email data set spanning seventeen county governments in North Carolina. By relying on public records requests, our data collection process is replicable. The resultant data set provides a micro-level view of manager-to-manager communication in these county governments, and a unique opportunity to study the relationship between a department manager's gender and the emails they send and receive.

We start by analyzing aggregate communication patterns from the numbers of emails sent and received by all department managers. In contrast to previous research, we find no strong evidence of the kinds of gender-homophilous or gender-heterophilous patterns that suggest gender bias. To unpack this finding, we investigate whether different discussion domains exhibit different communication patterns, using the sender's department as a proxy for discussion domain. By analyzing department-to-department communication patterns, we do find evidence of gender bias, with some departments exhibiting homophily and others exhibiting heterophily. Finally, to determine whether these patterns are driven by communication content rather than other factors correlated with man-

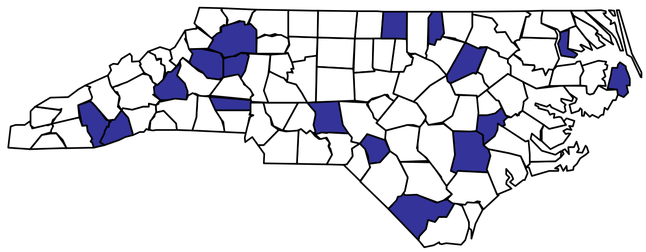


Fig. 1. The seventeen North Carolina counties used in our analysis (shaded).

Table 1. The numbers of male and female department managers for each county, along with the number of manager-to-manager emails sent.

	Manager Gender		
County	Male	Female	# Emails
Alexander	12	9	907
Caldwell	12	8	121
Chowan	12	11	2,027
Columbus	14	10	920
Dare	15	12	2,247
Duplin	13	14	1,914
Hoke	13	11	1,106
Jackson	18	6	1,499
Lenoir	15	5	560
Lincoln	15	7	573
McDowell	12	5	326
Montgomery	8	10	680
Nash	11	8	1,147
Person	12	9	1,491
Transylvania	16	4	1,857
Vance	10	8	185
Wilkes	15	2	303
Total	223	139	17,863

Table 2. Per-manager email statistics.

	Manager Gender	
	Male	Female
Average # emails sent	48.3	51
Average # recipients per email sent	1.45	1.43
Average # emails received	70.8	71.6

agers' departments, we use a recently developed latent variable model to identify topic-specific communication subnetworks for each county. We find differing degrees of gender homophily and heterophily for different topics of communication.

Data

We selected the state of North Carolina for our study because its public records laws explicitly mention email data and prevent counties from charging unreasonable fees for fulfilling requests. To construct our email data set, we issued public records requests to the one hundred North Carolina county governments. Our request to each county covered all emails sent and received by the department managers (e.g., health, finance, and elections) over a randomly selected period of three months between January and October in 2013. Twenty-three counties complied with our request, of which seventeen provided sufficient data for our analyses in an electronic format. Figure 1 indicates the seventeen counties. These counties are statistically indistinguishable from the other eighty-three counties in North Carolina along various demographic dimensions, including population, per-capita income, and percentage of the population that is white. In total, these seventeen counties produced over half a million emails, including 17,863 that were sent by a department manager to at least one other department manager in the same county (as well as other recipients, in some cases). We restricted our analyses to these manager-to-manager emails. To augment this data set, we also gathered information on the department affiliation and gender of each of the 362 managers represented in our data set. We provide some descriptive statistics for each county in table 1. Overall, almost 40% of the department managers are women, though there is significant variation across counties.

Descriptive Analysis

We begin our analysis of the relationship between a department manager's gender and the emails they send and receive, by looking for differences in the propensity for department managers to send emails to other managers of the same gender, and managers of the opposite gender. Table 2 provides some basic descriptive statistics including the average number emails sent and received by male and female department managers in our sample. On average, male and female department managers send and receive a comparable number of emails, and emails sent by male and female department managers have a similar number of recipients. Therefore, if we observe gender differences in email communication, they are unlikely to be driven by some innate difference in the propensity for male and female department managers to send or receive emails.

To test whether the gender of an email sender is related to the gender of its recipients, in aggregate, we construct a contingency table of email sending and receiving by gender (see table 3). We then perform a χ^2 test for independence between the rows and columns. The χ^2 test statistic we obtain indicates that the gender of an email sender and its recipients is not independent ($\chi^2 = 6.4, p = 0.011$). However, inspection of the contingency table indicates that the test statistic is actually driven by mild gender heterophily in communication. Furthermore, the gender differences we observe are substantively quite small, as both male and female department managers send emails to recipients of each gender in rough proportion to their overall representation in the sample (60% to men, 40% to women).

Table 3. Each cell records the number of times a department manager of gender X was included as a recipient of an email sent by a department manager of gender Y. Statistics provided are calculated for all counties combined. Note that each email may have more than one recipient.

		Recipient		
		Male	Female	Total
Sender	Male	9458	6120	15,578
	Female	6,330	3,833	10,163
	Total	15,788	9,953	

Domain-Dependent Gendered Communication. Within large organizations that perform heterogeneous functions, it is of little value to ask whether communication is gender biased in the aggregate. Broad patterns may be dominated by personal communications or mundane professional interactions that are inconsequential to the direction of the organization or the careers of employees. Examining domain specific patterns of communication will build upon limited existing findings, which indicate that gendering patterns depend upon the domain or context of communication. For example, Brass [5] finds a higher degree of male-male homophily in communication network domains that deal with long range strategic planning. We should also expect female managers to be relatively more likely to send heterophilous ties in communication network content partitions that deal with short and medium term coordination and planning, as research has found that female managers tend to preferentially communicate through formal channels [14] and favor gender-heterophilous instrumental connections [12]. The scope and scale of our data present an unprecedented opportunity for understanding whether and how gendering patterns vary with the domain of communication.

We take a two-pronged approach to understanding how the gendering of communication varies with the domain of communication within local government organizations. Our two empirical analyses are complimentary along two dimensions: (1) the use of within versus across-county variation; and (2) the operationalization of domain through email content versus the department affiliations of the managers.

Department Affiliation as Domain

Governments, and organizations more generally, spawn subordinate partitions in order to separate tasks according to the domain of organizational function. In the current analysis, we use the departmental affiliations of the managers in our data as proxy measures for the domain of organizational function about which managers communicate (or choose not to communicate) with each other. The use of department affiliation to measure communication domain raises a couple of distinct challenges in our strictly within-county e-mail data, which we solve by combining data across counties. First, aside from a few outlying cases, we only observe one or two managers with a given department affiliation in each county (e.g., one parks department manager). This means that we observe very little within-department communication in our data. Second, within each county, we only observe a handful of across-department dyads of a selected type (e.g., one parks/tax departments pair), which means that we cannot identify gendered patterns within counties that are specific to department pairs (i.e., domain-specific).

We solve the problems raised through using department affiliation to proxy domain by looking across counties for comparable department pairs. The department affiliations are comparable across counties, which means that for most department pairings, we observe a mix of male/male, female/female,

Table 4. Number of male and female managers for each department.

	Emergency	Manager	HR	Finance	IT	Health	Plan/Dev	Util/Waste	Tax	Parks/Rec	Soc_Serv	Transport	Info	Misc	Inspections	Maintenance	Library	Veterans	Seniors	Animal	Elections	Sheriff	Environment	Deeds	Extension
Male	29	15	3	5	11	6	17	15	11	9	8	8	2	5	13	5	3	5	2	9	2	16	9	6	8
Female	3	2	12	12	2	11	6	2	7	5	10	1	6	2	3	1	8	7	6	3	11	1	4	9	5
Total	32	17	15	17	13	17	23	17	18	14	18	9	8	7	16	6	11	12	8	12	13	17	13	15	13

female/male and male/female dyads. Using the department pairing to divide potential communication ties into domains, by looking across counties, we're able to observe how the intensity of communication within a given domain varies across different gender combinations. Table 4 gives the gender and department affiliation breakdown of managers in our data. Departments were hand coded into one of twenty-five different categories based on the title given in the county directory, to group departments that perform a similar function. Note that not all departments are represented in each county.

In our first approach to domain-specific analysis of gendering, we construct subsets of data that are specific to department pairings, but span all of the counties in which such pairings are observed. Consider, for example, the pairing of "Human Resources" and "Emergency Services" departments (HR and EMS, respectively). There is at least one HR and one EMS manager in fifteen counties. Across all counties we observe eight male/male pairs of HR/EMS managers, twenty-two mixed gender pairs and two female/female pairs. We construct department-pairing-specific datasets such as this for all of the 300 pairings that can be constructed using the twenty-five departments in our data.

For each of the 130 department-pairing datasets in which there is at least one of each gender pairing type, we fit two Poisson models (the base model and the gendered model) to the frequency of communication between managers. Let $y_{ij} \sim \text{Pois}(\exp(\eta_{ij}))$ be the number of e-mails from i to j. In the base model

$$\eta_{ij} = \beta_0,$$

and in the gendered model

$$\eta_{ij} = \beta_1 g_i g_j + \beta_2 g_i(1-g_j) + \beta_3(1-g_i)g_j + \beta_4(1-g_i)(1-g_j),$$

where g_i and g_i are indicators of the genders of the senders and recipients of e-mails, respectively, with males coded 0 and females 1. Since the gendered model reduces to the base model if $\beta_1 = \beta_2 = \ldots = \beta_4$, we use a likelihood-ratio test to evaluate whether the fit of the gendered model, in which separate rates are estimated for each directed gender pairing, is significantly better than the fit of the base (i.e., constant rate) model. We deem a domain, as represented by a pairing of departments, to exhibit gendered communication patterns if the gendered model fits statistically significantly better than the base model, and we use a p value of 0.05 as the threshold for determining statistical significance.

The gendered model fits better in approximately 70% (90 out of 130) of the domains. This result indicates that there is a substantial degree of within-domain gender bias. To summarize domain specific results in those domains in which we find significant gendering, we present lists of domains in which we find a consistent gendering pattern, for the six largest gendering patterns we identify. We take a gendering pattern to be a rank-ordering of the coefficients (β's) in the gendered model. The domains that exhibit the six most prevalent gendering patterns are presented in table 5. The most prevalent gendering pattern, depicted in the first column and first row of Table 5, is characterized by female-centric communication in which both females and males send communications to females at a higher rate than to males, and females send communication at a higher rate than do males. This gendering pattern is dominated by HR departments. The second most prevalent gendering pattern is also female-centric in that senders of both genders communicate more frequently with females than with males. Both the departments of Elections and Emergency Services are disproportionately represented in this pattern, with Elections appearing in pairs only within this pattern of gendering and Emergency Services appearing only once outside of this pattern. The two gendering patterns in the second row, and in the second column of the third row, are more male centric, with both females and males exhibiting bias in favor of male recipients. Soil and Water, County Extension, Sheriff, and Social Services departments appear either exclusively or disproportionately within these patterns. Looking across gendering patterns, we see that the forms of bias exhibited in communication with managers of some departments, including Information Technology, Health, Tax Administrator and County Manager, depend upon the affiliation of the other manager in the dyad, as these departments are relatively evenly spread across gendering patterns.

There are several limitations involved with using the department affiliations of managers to operationalize domain. First, we need to combine dyads across counties to build domain-specific datasets, which involves the questionable assumption that managers from comparably named departments, but very different counties (e.g., urban/rural, wealthy/poor, coastal/Appalachian), perform similar governing functions. Second, this approach does not make use of the textual content in the e-mails, which is likely relevant to understanding the domain of communication. Our second analytical approach addresses the limitations associated with using department affiliation to measure the domain of communication.

Email Content as Domain

In our second analytical approach we use the content of emails to determine the domain of communication. This is a natural and flexible way of accounting for the domain of communication, as emails about different domains will typically have different content. However, the challenge with using the raw text to classify emails according to domain is that language use does not vary in a perfectly predictable manner across domains. As such, we use a statistical model that has been developed by Denny et al. [7] for application to text-valued interpersonal communication data, in order to infer domains from patterns of word co-occurence and sender-receiver interaction. The modeling framework we adopt represents a joint model of email content and content-specific network structure, which captures the content-conditional relationship between the gender of an email sender and the gender of its recipients.

Table 5. Domains in which there is significantly gendered rates of communication grouped by gendering pattern.

FF>FM>MF>MM	FF>MF>FM>MM
HR & Health	Information Technology & Health
HR & Information Technology	HR & Emergency Services
HR & County Manager	Library & County Manager
Planning & HR	Register of Deeds & Information Technology
Register of Deeds & HR	Parks and Recreation & Health
Parks and Recreation & HR	Parks and Recreation & County Manager
Finance & HR	Finance & County Manager
Finance & Parks and Recreation	Finance & Planning
Social Services & HR	Veteran Services & Information Technology
Solid Waste and Recycling & HR	Elections & Emergency Services
Tax Administrator & HR	Elections & Information Technology
Tax Administrator & Library	Elections & County Manager
Tax Administrator & Finance	Animal Control & Emergency Services
Code Enforcement & HR	Soil and Water & Planning
Animal Control & HR	
MM>FM>MF>FF	**FM>MM>MF>FF**
Planning & Information Technology	Social Services & County Manager
Solid Waste and Recycling & Health	Sheriff & Social Services
Sheriff & Health	Tax Administrator & Parks and Recreation
Tax Administrator & Planning	Tax Administrator & Veteran Services
Tax Administrator & Social Services	Animal Control & Tax Administrator
Code Enforcement & Tax Administrator	Animal Control & Code Enforcement
Animal Control & Finance	Soil and Water & HR
Soil and Water & Health	Soil and Water & Finance
Soil and Water & Solid Waste and Recycling	
FF>MM>FM>MF	**MM>MF>FM>FF**
County Manager & Health	Parks and Recreation & Planning
Planning & County Manager	Social Services & Planning
Finance & Emergency Services	Veteran Services & Register of Deeds
Finance & Health	Airport & Health
Social Services & Health	Airport & Tax Administrator
Social Services & Information Technology	County Extension & Planning
Social Services & Parks and Recreation	County Extension & Parks and Recreation

This approach compliments the department-based analysis in two ways.

A Model of Email Content. Denny et al.'s model jointly accounts for the structure and content of an observed email network by combining ideas from latent space network modeling [11] with ideas from statistical topic modeling [4]. This model treats the sender, recipients, and contents of each email as observed, and simultaneously infers latent topics of communication and content-based gender mixing patterns.

Denny et al.'s model is a generative model; that is, it implies a particular probabilistic generative process, by which a corpus of emails could theoretically have been generated.

This generative process starts by generating the global (corpus-wide) variables. There are two main types of global variables: those that describe the topics people talk about and those that descibe how people interact (interaction patterns). The former are a set of T topics. Each topic $\boldsymbol{\phi}^{(t)}$ is a discrete distribution over V word types. The latter are a set of C interaction patterns. Each interaction pattern consists of an intercept $b^{(c)} \in \mathbb{R}$, a coefficient vector $\boldsymbol{\gamma}^{(c)} \in \mathbb{R}^P$, and a set of A positions $\{\boldsymbol{s}_a^{(c)} \in \mathbb{R}^K\}_{a=1}^A$—one for each person. Each sender–recipient pair is also associated with an observed P-dimensional vector of covariates $\boldsymbol{x}^{(ar)}$; these covariates are not generated, however. Together these variables specific the (pattern-specific) probability of sender $a \in \{1, \ldots, A\}$ emailing recipient $r \neq a$: $p_{ar}^{(c)} = \sigma(b^{(c)} + {\boldsymbol{\gamma}^{(c)}}^\top \boldsymbol{x}^{(ar)} - ||\boldsymbol{s}_a^{(c)} - \boldsymbol{s}_r^{(c)}||)$.

The topics and interaction patterns are linked via a set of T cateogorical variables. Each variable l_t associates the corresponding topic with a single interaction pattern that best describes how people interact when talking about that topic.

Next, the generative process generates the local (email-specific) variables. There are D emails. Each email's sender $a^{(d)} \in \{1, \ldots, A\}$ and length $N^{(d)} \in \mathbb{N}^0$ are not generated. First, each email is associated with a distribution $\boldsymbol{\theta}^{(d)}$ over the T topics. Each token $w_n^{(d)}$ in the email is generated by first drawing a topic $z_n^{(d)}$ from this distribution and then drawing a word type from the topic's discrete distribution $\boldsymbol{\phi}^{(z_n^{(d)})}$. Having generated the email's contents, the generative process then proceeds by generating its recipients. For each possible recipient $r \neq a^{(d)}$, a binary variable $y_r^{(d)}$ is generated indicating whether or not the email is sent to that recipient. This variable is drawn from a Bernoulli distribution, parameterized by a mixture of pattern-specific probabilities: $\sum_{c=1}^C \frac{\pi^{(c)}}{\sum_{c=1}^C \pi^{(c)}} p_{a^{(d)}r}^{(c)}$. The unnormalized mixing weight $\pi^{(c)}$ for pattern c is the number of tokens associated with (a topic associated with) pattern c in that document. As a result, the

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

[2] We use 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 LDA, and provides good performance.

[3] male–male, male–female, female–male, and female–female

for $t = 1$ to T **do**
  draw $\boldsymbol{\phi}^{(t)} \sim \text{Dir}(\beta,)$
end for
for $c = 1$ to C **do**
  draw $b^{(c)} \sim \mathcal{N}(\mu, \sigma_1^2)$
  draw $\boldsymbol{\gamma}^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \, \mathrm{I}_P)$
  for $a = 1$ to A **do**
    draw $\boldsymbol{s}_a^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \, \mathrm{I}_K)$
  end for
  for $a = 1$ to A **do**
    for $r = 1$ to A **do**
      if $r \neq a$ **then**
        set $p_{ar}^{(c)} = \sigma(b^{(c)} + \boldsymbol{\gamma}^{(c)\top} \boldsymbol{x}^{(ar)} - ||\boldsymbol{s}_a^{(c)} - \boldsymbol{s}_r^{(c)}||)$
      else
        set $p_{ar}^{(c)} = 0$
      end if
    end for
  end for
end for
for $t = 1$ to T **do**
  draw $l_t \sim \text{Unif}(1, C)$
end for
for $d = 1$ to D **do**
  draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha,)$
  set $\bar{N}^{(d)} = \max(1, N^{(d)})$
  for $n = 1$ to $\bar{N}^{(d)}$ **do**
    draw $z_n^{(d)} \sim \boldsymbol{\theta}^{(d)}$
    if $N^{(d)} \neq 0$ **then**
      draw $w_n^{(d)} \sim \boldsymbol{\phi}^{(z_n^{(d)})}$
    end if
  end for
  for $t = 1$ to T **do**
    set $\bar{N}^{(t|d)} = \sum_{n=1}^{\bar{N}^{(d)}} \delta(z_n^{(d)} = t)$
  end for
  for $r = 1$ to A **do**
    draw $y_r^{(d)} \sim \text{Bern}(\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(l_t)})$
  end for
end for

Fig. 2. Generative process for Denny et al.'s model.

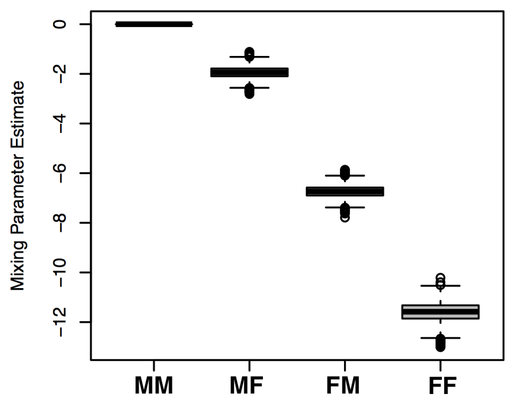
email's recipients depend on the topics expressed in that email and the interaction patterns associated with those topics.

This generative process implies a particular factorization of the joint distribution over $\Phi = \{\boldsymbol{\phi}^{(t)}\}_{t=1}^T$, $\mathcal{B} = \{b^{(c)}\}_{c=1}^C$, $\Gamma = \{\boldsymbol{\gamma}^{(c)}\}_{c=1}^C$, $\mathcal{S} = \{\{\boldsymbol{s}_a^{(c)}\}\}_{c=1}^C$, $\mathcal{L} = \{l_t\}_{t=1}^T$, $\Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^D$, $\mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^D$, $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^D$, and $\mathcal{Y} = \{\boldsymbol{y}^{(d)}\}_{d=1}^D$ given $\mathcal{X} = \{\{\boldsymbol{x}^{(ar)}\}_{r=1}^A\}_{a=1}^A$, $\mathcal{A} = \{a^{(d)}\}_{d=1}^D$, and $\mathcal{N} = \{N^{(d)}\}_{d=1}^D$.

The complete generative process is provided in figure **??**

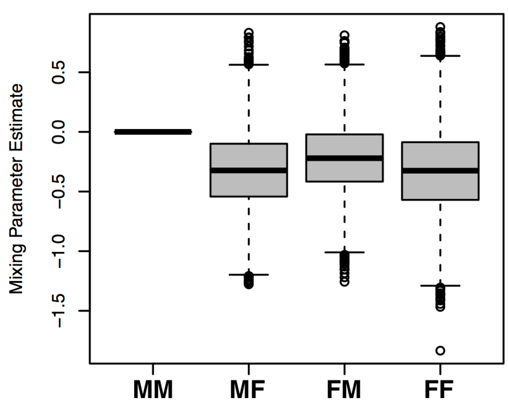
Inference. For real-world email networks, we must invert the generative process described in the previous section to infer plausible values for the latent variables Φ, $\mathcal{B}$, Γ, $\mathcal{S}$, $\mathcal{L}$, Θ, and $\mathcal{Z}$. Denny et al. achieve this goal by integrating out Φ and Θ and then drawing samples from the posterior distribution over $\mathcal{B}$, Γ, $\mathcal{S}$, $\mathcal{L}$, and $\mathcal{Z}$ given $\mathcal{W}$, $\mathcal{Y}$, $\mathcal{X}$, and $\mathcal{A}$. Specifically, they define a Metropolis-within-Gibbs algorithm, in which each iteration involves sequentially resampling the value of each $z_n^{(d)}$ variable from its conditional posterior distribution, sequentially resampling the value of each l_t variable similarly, and then jointly sampling the values of $\mathcal{B}$, Γ, and $\mathcal{S}$ using the Metropolis algorithm. This procedure is in algorithm **??**.[1].

We apply our model separately to the email data from each county and then pool our model results. To perform inference on real data, we must first select a number of model hyper-parameters. In particular, we must select the number topics, number of clusters, and topic model hyper-parameters[2] to be used by our model, which we hold constant across counties. We choose to include manager gender as the only covariate in our specification, and include the full complment of gender mixing parameters in our model[3]. In addition, we fix the male-male mixing parameter at zero for our analysis, to aid in directly interpreting the other gender mixing parameters. We select 40 topics and 4 clusters, to provide reasonable granularity in capturing variation in the content of communication, while improving the interpretability of the latent space model results by constraining the number of possible patterns



Topic top words
will, track, winds, system, forecast, atlantic, east
storm, sandy, high, coastal, tides, night, hurricane
status, update, boat, today, weather, people, ago
board, meeting, planning, seafood, will, hearing, public
box, planning, director, manteo, permit, building, collector

Fig. 3. Mixing parameter estimates and topic top words for the disaster response topic-cluster in Dare county. Topics are presented (one per line) in decreasing order of use within the topic-cluster, as are words within each topic.



Topic top words
junk, box, summary, emails, will, login, email
description, cell, director, works, public, fax, office
pharmacist, good, will, send, morning, schedule, day
director, communications, emergency, central, status, check
box, director, fax, work, address, will, payroll

Fig. 4. Mixing parameter estimates and topic top words for the IT and health related topic-cluster in Dare county. Topics are presented (one per line) in decreasing order of use within the topic-cluster, as are words within each topic.

Table 6. Most common completely unique words in topic clusters associated with three gender mixing patterns.

FM > MF > FF > MM	MM > FM > MF > FF	FF > FM > MF > MM
junk, summary, emails, fire, documents, accompanying, economic, service, marshal, prohibited, meter, messages, notify, login, delete, long, personal, visit, parks	**church, suite,** operations, jail, fort, project, pool, april, full, cashiers, book, ext, books, cpa, including, story, works, march, phase, lines, staff, court, west, dual, thursday, force, free, executive, reading, tourism, authority, sheriff, correspondence, tuesday, communications, copy, morning, buildings, taps, enp, renewal, corner	**fyi,** debt, learn, great, actions, day, inspire, supply, dream, animal, form, item, shelter, audit, funds, refunding, site, fund, requests, timesheet, today, media, cprp
MF > FM > FF > MM	**FF > MF > FM > MM**	**MM > FF > MF > FM**
payment, notification, munis, computer, program, position, excellent, cemetery, fiscal, requirements, lane, company, planner/section, user, entered, e/s, customer, commodity, manager/redi, opportunity, applicants, accessed, ability, generated	transportation, administrator, worker, comp, ipad, change, call, check, legion, well, utilities, policy, dss, increase, matt, reminder, record, salary, send, cost, attorney, facilities, hey, airport, response, balance, help, loss, defensive, monday, unaccounted, expenditure	funding, interim, john, requested, notified, tel, seal, mph, college, addressees, required, entity, document, july

Table 7. Most common exclusive words in topic clusters associated with three gender mixing patterns.

FM > MF > FF > MM	MM > FM > MF > FF	FF > FM > MF > MM
cell, employees, message, class, junk, summary, benefits, confidential, plan, error, emails, fire, survey, state, main, report, documents, system, accompanying, economic, insurance, recipient, service, marshal, enrollment, prohibited	**received, description, law, mail, planning, intended, records, church, suite, subject, third, address, tax, parties, water,** marion, operations, electronic, jail, fort, project, pool, april, full, center, insurance, cashiers, book	**east, fyi,** password, insurance, washington, contact, zee, main, debt, learn, great, actions, day, inspire, disclosure, supply, dream, animal, form, item, shelter, audit, funds, refunding, site
MF > FM > FF > MM	**FF > MF > FM > MM**	**MM > FF > MF > FM**
requisition, payment, approval, notification, munis, class, computer, program, general, position, system, plan, excellent, benefits, chuck, pending, cemetery, week, fiscal, requirements, lane, company, planner/section, user, entered, e/s, customer, commodity	**budget, read, work, request, emergency,** password, transportation, electronic, administrator, worker, comp, ipad, officer, change, call, check, center, legion, well, utilities, report, review, policy, dss, increase,	**services,** class, benefits, funding, recipient, attachments, plan, questions, notice, requisition, pending, survey, interim, john, enrollment, error, requested, approval, notified, disclosed, tel, insurance, seal, mph, electronic, officer

of communication. This choice has a practical advantage of ensuring that enough data will be available to fit each clusters latent space with reasonably low uncertainty in the parameter estimates.

To perform inference, we must also select a number of iterations for our MCMC sampler. In the first step of inference, we alternate between one iteration of Gibbs sampling for the LDA parameters, and 1,000 iterations of Metropolis Hastings sampling for the LSM 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 LDA model log likelihood for all counties. We then ran the LSM component of our model for an additional 10,000,000 iterations, holding the LDA parameters fixed, to ensure that all latent space parameter estimates had converged.

Analysis. Our model produces three key outputs that will be useful to an analysis of content specific patterns of gender mixing in communication. First, it infers a set of (in this case) 40 topics of communication for each county. These topics are distributions over words, and are commonly summarized by listing the words that are most frequently assigned to them. For example, a law enforcement topic might have the following top 10 words: *safety, management, understanding, training, law, enforcement, local, basics, workplace, corrections*. Our model also associates each topic with one of (in this case) 4 clusters, based on a common pattern of communication. Finally, a set of gender mixing parameters is inferred for each cluster. Therefore, we can use the mixing parameter estimates and associated topic top-words inferred by our model as ingredients in an analysis that is similar to the one presented in the previous section.

Before diving into this analysis, it will be useful to examine a few examples of our model output. Our data collection window happend to overlap with Hurricane Sandy (October, 2013), and one of the counties in our sample (Dare county) happens to be located on the outer banks, so we might expect one domain of communication in this county to be disaster preparation and response. As illustrated in figure 3, our model infers a disaster response topic-cluster. We can see from the mixing parameter plot that this topic-cluster is strongly male-centric, which makes sense given that the emergency managers and county manager are all male in this county. Importantly, our model also assigns some topics that follow a similar pattern of communication, but are not disaster related, to this cluster – such as topics related to costal and fisheries management. In contrast to the above example, other topic-clusters display no discernable gender bias. This is well illustrated by an IT and health related topic-cluster presented in figure 4 from Hoke county.

Our analysis goal is to construct a table similar to table 5, but now we will relate rank-orderings of inferred gender mixing parameters to words that appear most frequently in the topics associated with them. To do this, we first generate a list of the twenty words that appear most frequently in the topics associated with each cluster, in each county. This yields us 16 counties $\times$ 4 clusters = 64 observations where top words are related to mixing parameters. We then rank-order the mixing parameter estimates in each cluster and choose to examine three patterns of rank orderings. The first of these is the case where none of the inferred mixing parameters are significantly different from the male/male base-case (MM = MF = FM = FF). The second case is male-centric communication (MM > MF/FM > FF), and the third case is female-centric communication (FF > MF/FM > MM). We find that 11 of the 64 clusters are associated with each of these patterns. We then combine the top 20 words associated with each cluster to produce a list of the words that appear most frequently in topics associated with clusters that follow each of these three gender mixing patterns.

What we are particularly interested in are the words that distinguish between each of these communication patterns. We therefore filter the lists of top words associated with each pattern to only those words that appear one or fewer times in either of the other lists. The reason for this is that we do not wish to exclude words which appear very frequently in topics associated with one pattern but appear only once in topics associated with the others. The results of this filtering are presented in table 7. As we can see, there is significant variation in the distinctive words across the three patterns of communication. For example, the top words associated with the MM = MF = FM = FF pattern seem to be mostly related to daily operations and legal matters, while the top words associated with the MM > MF/FM > FF pattern are mostly related to planning, education, and development. While it is not unexpected that finance and budget

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

References

1. Susan M Adams, Atul Gupta, Dominique M Haughton, and John D Leeth. Gender differences in ceo compensation: evidence from the usa. *Women in Management Review*, 22(3):208–224, 2007.
2. James Albrecht, Anders Björklund, and Susan Vroman. Is there a glass ceiling in Sweden? *Journal of Labor Economics*, 21(1):145–177, 2003.
3. William T. Bielby and James N. Baron. Men and Women at Work : Sex Segregation and Statistical Discrimination. *American Journal of Sociology*, 91(4):759–799, 1986.
4. DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
5. Daniel J. Brass. Men's and Women's Networks: A Study of Interaction Patterns and Influence in an Organization. *Academy of Management Journal*, 28(2):327–343, 1985.
6. Emilio J Castilla. Gender, race, and meritocracy in organizational careers. In *Academy of Management Proceedings*, volume 2005, pages G1–G6. Academy of Management, 2005.
7. Matthew J. Denny, Hanna Wallach, and Bruce A. Desmarais. Hierarchical topic-partitioned multinetwork embeddings: Modeling social structure and content across a sample of communication networks. In *International Conference on Computational Social Science , Helsinki*, 2015.
8. Colin Duncan and Wendy Loretto. Never the Right Age? Gender and Age-Based Discrimination in Employment. *Gender, Work & Organization*, 11(1):95–115, 2004.
9. Kim M Elsesser and Janet Lever. Does gender bias against female leaders persist? quantitative and qualitative data from a large-scale survey. *Human Relations*, 64(12):1555–1578, 2011.
10. Robin J Ely and Debra E Meyerson. Theories of gender in organizations: A new approach to organizational analysis and change. *Research in organizational behavior*, 22:103–151, 2000.
11. 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.
12. Herminia Ibarra. Homophily and Differential Returns : Sex Differences in Network Structure and Access in an Advertising Firm. *Administrative Sceince Quarterly*, 37:422–447, 1992.
13. Peter Krafft, Juston Moore, Bruce A Desmarais, and Hanna Wallach. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems Twenty-Five*, 2012.
14. Belle Rose Ragins and Eric Sundstrom. Gender and Power in Organizations: A Longitudinal Perspective. *Psychological Bulletin*, 105(1):51–88, 1989.