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 shaping interpersonal communication in local government organizations. Our central contribution is the identification of the domains of organizational function in which we observe specific patterns of gender-dependent communication. We analyze email corpora from seventeen county governments in North Carolina, which cover all aspects of local government organizational function. To investigate this finding, we therefore analyze department-to-department communication patterns. We separately treat both the department affiliations of county employees and the textual content of emails as measures of the domains of communication. Applying both dyadic regression and a recently developed latent variable model to identify topic specific communication subnetworks for each county, we find robust patterns in terms of how interaction depends upon gender across discussion domains. Our results agree with recent findings in the literature indicating that females are more active in small groups when the groups are majority female.

Results are preliminary and may change, please contact authors before citing or re-distributing (mdenny@psu.edu)

Gender in Organizations

Researchers have observed and documented gender bias in organizational processes in both the public and private sectors. Women have been found to receive lower pay, hold less prestigious positions, have reduced opportunities for advancement, and are excluded from decision-making coalitions [5, 3, 14, 2, 9]. In contrast, organizations that aspire to a just, efficient, and sustainable culture strive to provide men and women with equal treatment in the workplace [11]. 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 on day-to-day organizational function means that most research is based on either small-scale ethnographic data, self-reports, or aggregate employee-level official data [e.g., 6, 1, 10]. Since these data sources are restricted in scope, can be biased by subjective assessments, or aggregate over micro-level functioning, their use in understanding gender bias is limited.

In this paper, we address limitations of prior research through the study of internal email corpora, which constitutes primary-source data on day-to-day interactions. We focus specifically on local government organizations and seek to understand the role of gender in shaping communication patterns. With the increasing use of electronic communication in the workplace, and the rise of transparency initiatives within government, internal email corpora in many government organizations can be accessed through public records requests. We draw upon this resource to construct an email data set spanning seventeen county governments in North Carolina. 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 leverage both the breadth and depth of information represented in our data to identify patterns of gender-dependent interaction that are specific to different domains of organizational function. First, we use the breadth of coverage in our data—the fact that we observe several managers with comparable departmental affiliations, to analyze how patterns of gender-dependent interaction differ across domains, as measured by department affiliation. Second, drawing upon the qualitative depth represented in email text, we use a recently developed latent variable model to identify topic specific communication subnetworks for each county. Among other patterns, in both approaches to analyzing our data, we find that female-centric communication patterns, which are those in which both males and females are statistically more likely to send emails to females than to males, are observed in domains represented by female majorities.

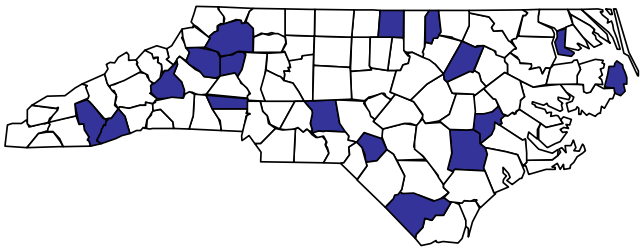


Fig. 1. The seventeen North Carolina counties used in our analysis (shaded).

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

[1]Our validation experiments indicate a 100 percent agreement between human identification of the text that should remain in our emails, as well as e-mail metadata, and our automated text extraction approach. A separate technical report detailing these efforts is available upon request.

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 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 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.

We also took a number of steps to prepare the text of these emails for analysis. We began by removing any quoted text using a series of heuristics and indicators provided by the email client which was used to create them (for example Microsoft Outlook email files share a common delineator for quoted text). We then ran the text of the emails through the Stanford CoreNLP text processing libraries [16] and named entity recognizer [12] to tokenize the text and remove any resulting tokens that were classified as the name of a person, before removing a list of common and domain specific stopwords. We then preformed a series of validation experiments using the raw email data to verify that the final text we extracted accurately reflected the original email text.[1] Descriptive statistics detailing the number of unique words and total number of tokens in each county email corpus (after preprocessing) is provided in table 2.

Table 2. County email corpus descriptive statistics. The number of tokens is the sum of all tokens (after filtering) in all emails.

	# Emails	# Unique Words	# Tokens
Alexander	907	1595	15437
Caldwell	121	94	2870
Chowan	2027	3038	39635
Columbus	920	1326	13763
Dare	2247	2907	38174
Duplin	1914	2380	83162
Hoke	1106	1397	19409
Jackson	1499	2493	34290
Lenoir	560	1340	19780
Lincoln	573	1458	13754
McDowell	326	1178	8863
Montgomery	679	1137	11166
Nash	1147	2214	31629
Person	1491	2618	29153
Transylvania	1857	3157	57431
Vance	185	620	4070
Wilkes	303	510	5594

Table 3. 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

Table 4. 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	

Aggregate Patterns by Gender

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 3 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.

In the interest of being concise, we use the term *gendered* to refer to patterns of communication where we observe statistically significant variation between the propensity for men to send emails to men, men to send emails to women, women to send emails to men, or women to send emails to women. To test whether aggregate communication flows in our data are gendered, we construct a contingency table of email sending and receiving by gender (see table 4). 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 a modest degree of gender heterophily – a pattern by which cross-gender interactions are more likely than within-gender interactions – 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).

Domain-Dependent Gendered Communication. Within large organizations that perform heterogeneous functions, there is little reason to expect aggregate communication patterns and biases to apply to the different functions and sub-areas of the organizations' operations. Aggregate 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. we use the term *domain* to refer to an organizational function with which a communication instance is concerned (e.g., long-range strategic planning, human resource management, public relations). Examining domain specific patterns of communication will build upon limited existing findings, which indicate that gendered patterns depend upon the domain or context of communication. For example, Brass [5] finds a higher degree of male-male homophily – the tendency of males to communicate with males at a higher rate than with females – in communication domains related to long range strategic planning. We should also expect female managers to be relatively more likely to send heterophilous ties in communication domains that deal with short and medium term coordination and planning, as research has found that female managers tend to preferentially communicate through formal channels [17] and favor gender-heterophilous instrumental connections [14]. The scope and scale of our data present an unprecedented opportunity for understanding whether and how gendering patterns vary with the domain of communication.

Measuring the domain of communication, given detailed data on communication content, is more challenging than it may at first appear. A single communication instance may be concerned with more than one domain. For example, a discussion thread concerned with composing a job advertisement may involve both the human resource management and public relations domains. Furthermore, individuals may use the same language to discuss different domains. For example, one employee may commend another on the results of a project, exclaiming that the project leader deserves a promotion. Such a communication may concern human resource management, but it may be a show of enthusiasm concerning the domain with which the project is concerned, and have no real bearing on personnel decisions. Given the importance of domain in understanding the structure of organizational communication networks, and the ambiguities that arise in measuring the domain of communication, we take a two-pronged empirical approach to the analysis of domain-dependent communication patterns. 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 units 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 affiliations to measure communication domain raises a couple of distinct challenges in our strictly within-county email 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 finance 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 health/tax departments pair), which means that we cannot identify gendered patterns within counties that are specific to department pairs (i.e., domain specific).

We address the challenges raised through using department affiliations 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 and female/male 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 5 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, 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, and at least ten pairs total (resulting in at least twenty directed dyadic observations), we fit two Poisson models, which we refer to as 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 emails 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 emails, 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 define a gendering pattern as the 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 6, where prevalence is measured as the number of domains that exhibit the respective pattern. The most prevalent gendering pattern, depicted in the first column and first row of Table 6, 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

Table 5. 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

Table 6. Domains in which we find gendered 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
Inspections & HR	Environment & 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
Inspections & Tax Administrator	Animal Control & Inspections
Animal Control & Finance	Environment & HR
Environment & Health	Environment & Finance
Environment & 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	Transport & Health
Social Services & Health	Transport & Tax Administrator
Social Services & Information Technology	County Extension & Planning
Social Services & Parks and Recreation	County Extension & Parks and Recreation

than do males. The second most prevalent gendering pattern is also female-centric in that senders of both genders communicate more frequently with females than with males. The more male-centric communication patterns include those in the second row and third row/second column of Table 6. The pattern represented in the first column and third row is homophilous, with both males and females communicating at high rates within gender. The departments of Human Resources, Finance, Elections and Emergency Services are disproportionately represented in the more female-centric patterns. The fact that departments of Human Resources, Finance and Elections are all represented by female majorities, is consistent with recent findings that females are more actively engaged in communication when they represent a majority in a group [7]. Environment, County Extension, Sheriff, and Social Services departments appear either exclusively or disproportionately within the more male-centric 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 emails, 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 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. [8] for application to text-valued interpersonal communication data, in order to infer domains from patterns of word co-occurrence 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. Unlike in the department-based analysis, we conduct the content-based analysis within counties, inferring a completely separate model for each county. This structure compliments the department-based approach, and avoids the problem of questionable comparability across counties.

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 [13] 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, extending Krafft et al.'s topic-partitioned multinetwork embeddings model [15].

Denny et al.'s model is derived from 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 describe 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 P-dimensional 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 fixed covariates $\boldsymbol{x}^{(ar)}$. Together these variables specify 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 categorical variables. Each variable l_t associates the corresponding topic with a single interaction pattern that 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 fixed. 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: $\sum_{c=1}^{C} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(c)}$. The normalized mixing weight $\frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}}$ for interaction pattern c is the proportion of tokens in that email associated with (a

for $t = 1$ to T **do**
  draw $\boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta, \boldsymbol{n})$
end for
for $c = 1$ to C **do**
  draw $b^{(c)} \sim \mathcal{N}(0, \sigma_1^2)$
  draw $\boldsymbol{\gamma}^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathrm{I}_P)$
  for $a = 1$ to A **do**
    draw $\boldsymbol{s}_a^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_3^2 \mathrm{I}_K)$
  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 $l_t \sim \mathrm{Unif}(1, C)$
end for
for $d = 1$ to D **do**
  draw $\boldsymbol{\theta}^{(d)} \sim \mathrm{Dir}(\alpha, \boldsymbol{m})$
  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 $c = 1$ to C **do**
    set $\bar{N}^{(c|d)} = \sum_{n=1}^{\bar{N}^{(d)}} \delta(l_{z_n^{(d)}} = c)$
  end for
  for $r = 1$ to A **do**

    draw $y_r^{(d)} \sim \mathrm{Bern}(\sum_{c=1}^{C} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(c)})$
  end for
end for

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

for $i = 1$ to I **do**
  for $d = 1$ to D **do**
    for $n = 1$ to $N^{(d)}$ **do**
      $z_n^{(d)} \sim P(z_n^{(d)} \mid \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A})$
    end for
  end for
  for $t = 1$ to T **do**
    $l_t \sim P(l_t \mid \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}_{\setminus t}, \mathcal{Z}, \mathcal{X}, \mathcal{A})$
  end for
  $\mathcal{B}, \Gamma, \mathcal{S} \sim P(\mathcal{B}, \Gamma, \mathcal{S} \mid \mathcal{L}, \mathcal{Z}, \mathcal{Y}, \mathcal{X}, \mathcal{A})$
end for

Fig. 3. Block Metropolis-within-Gibbs inference algorithm for Denny et al's model.

topic associated with) pattern c. As a result, the email's recipients depend on the topics expressed within it and, in turn, on 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 2.

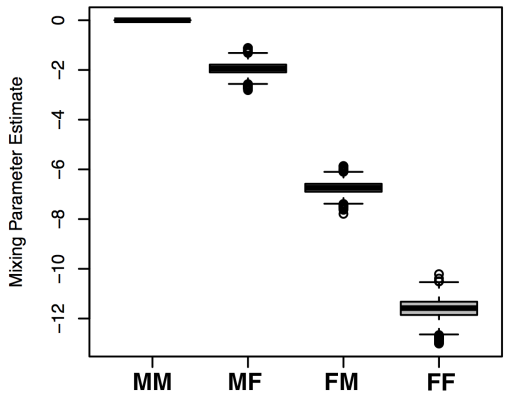
Inference. For real-world email networks, we must invert the generative process described in the previous section to infer plausible values for the latent variables Φ, $\mathcal{B}$, Γ, $\mathcal{S}$, $\mathcal{L}$, Θ, and $\mathcal{Z}$. Denny et al. achieve this goal by integrating out Φ and Θ and then 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 outlined in figure 3.

We applied Denny et al.'s model separately to the email data from each county and then aggregated the model results. We used uniform base measures for the Dirichlet priors and set $\alpha = 1$ and $\beta = 0.01V$, where V is the length of the vocabulary for each county. We also set $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 5$. We used forty topics and four clusters, to provide reasonable granularity in capturing variation in content, while making sure that the interaction patterns were interpretable and did not exhibit redundancy. We used the same values for all counties. Since our goal was to study gendered communication, we used four binary gender mixing covariates (i.e., MM, MF, FM, and FF). To ensure identifiability and interpretability of the coefficients Γ, we fixed the coefficient for the MM covariate to zero.

For each county, we ran Denny et al's inference algorithm for 4,000 iterations. This was sufficient to reach convergence (indicated by Geweke statistics) for all counties. To ensure mixing of Denny et al's inference algorithm, we draw 1,000 samples of $\mathcal{B}$, Γ, $\mathcal{S}$ during each iteration (line 10). After the 4,000 iterations, we fixed the values of $\mathcal{Z}$ and $\mathcal{L}$ and resampled the remaining variables for an additional 10,000,000 iterations.

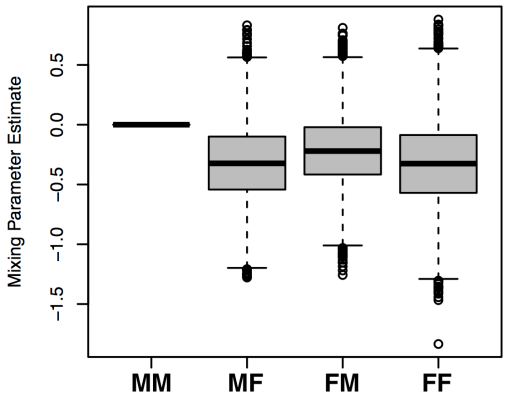
Results. Denny et al.'s model produces three key outputs that will be useful to an analysis of content specific patterns of gender mixing in communication. First, it infers a set of forty topics of communication for each county. These topics are distributions over words, and are effectively summarized by listing the words that have the highest posterior probability of being assigned to that topic. For example, a law enforcement topic might have the following top ten words: *safety, police, law, training, jail, enforcement, local, firearm, crime, corrections.* Second, Denny et al.'s model also associates each topic with one of four interaction patterns. Third, a set of gender mixing parameters is inferred for each interaction pattern. We use the mixing parameter estimates and associated topic top-words inferred by the model to summarize our results in a fashion that is analogous to our analysis using department affiliations as proxy for the domains of communication.

Before presenting our comprehensive results, we provide a few examples of model output. We summarize an interaction pattern with both a box plot of the posterior samples of regression coefficients corresponding to sender-receiver gender combinations and a table reporting the top words in topics associated with the respective interaction patterns as well as a hand-coded label of each topic. The mixing parameter two-letter acronyms are given by the letter corresponding to the gender of the sender and the gender of the prospective recipient, respectively. Note that the MM coefficient is fixed at zero such that the other mixing parameters should be interpreted as the difference in the log odds of a sender of a respective gender adding a recipient to an email given the gender of the respective recipient and the log odds of a male sender adding



Coding	Topic top words
Sandy	will, track, winds, system, forecast
Sandy	storm, sandy, high, coastal, tides, night
Harbor	status, update, boat, today, weather
Planning	board, meeting, planning, seafood, hearing
Storm	box, planning, director, permit, building

Fig. 4. Mixing parameter estimates and topic top words for the hurricane related interaction pattern in Dare county. Topics are presented (one per line) in decreasing order of use within the interaction pattern, as are words within each topic. The hand coding procedure is discussed below.



Coding	Topic top words
IT	junk, box, summary, emails, login
Public Works	description, cell, director, works, public
Health	pharmacist, good, morning, schedule
Emergency	communications, emergency, status
HR	box, director, fax, work, address, payroll

Fig. 5. Mixing parameter estimates and topic top words for the selected interaction pattern in Dare county. Topics are presented (one per line) in decreasing order of use within the interaction pattern, as are words within each topic. The hand coding procedure is discussed below.

[2] We compared this approach to simply weighting documents based on the number of tokens assigned to the topic, or the proportion of tokens assigned to the topic, and we found that it yields more interpretable emails.

a recipient to an email given that the prospective recipient is a male.

Our data collection window happened to overlap with Hurricane Sandy (October, 2013), and one of the counties in our sample (Dare county) is located on the coast, so we would expect there to be some hurricane related topics in our model output. As illustrated in figure 4, our model infers an interaction pattern in which a number of the associated topics are related to Hurricane Sandy. We can see from the mixing parameter plot that this interaction pattern is strongly male-centric. In contrast to the above example, other interaction patterns display no discernible gender bias. This is well illustrated by the interaction pattern presented in figure 5 from Hoke county.

For our analysis, we seek to construct a table similar to table 6, but now we relate rank-orderings of inferred gender-mixing parameters to topics associated with them. To do this, we first rank-order the mixing parameter estimates in each interaction pattern, for each county. We then discard those interaction pattern observations where we fail to reject the null hypothesis that all mixing parameters are equal. We then examine the topics from those interaction patterns that display one of the six most prevalent rank orderings of gender mixing parameters, as determined by the total number of tokens assigned to all topics in all interaction patterns associated with a particular rank ordering. Table 7 displays the three topics tow which the largest number of tokens are assigned, from each of the interaction patterns associated with each of the six most prevalent gender mixing parameter rank orderings. The gender mixing patterns are displayed in descending order of the total number of tokens that were assigned to all topics associated with them (from top left to bottom right, by rows). Note that in this analysis, we only make use of data from sixteen counties, as the email data from Caldwell county did not contain a sufficient number of tokens per email (after preprocessing) to facilitate interpretation of model results.

In addition to displaying the top words associated with each topic, we assigned a hand-coded label to each topic using the following procedure. First, for each topic (which is specific to a county) we calculated a weight for each email based on the representation of that topic in the email using the following weighting scheme.

$$\omega_t^{(d)} = \frac{\left(N_t^{(d)}\right)^2}{N^{(d)}} \qquad [1]$$

In words, the weight we assign each document is equal to the proportion of tokens in that email that are assigned to the topic, times the number of tokens in that email assigned to the topic. This gives the highest weight to long emails that contain a high proportion of tokens assigned to the topic^2. We then selected the ten highest weight emails and read them, before assigning a label to the topic. This hand coding was performed by one member of our team, and a label was selected using the following criterion: the coder selected a word or phrase that best summarized the common thread between the emails, unless no common thread was detected, in which case the topic was labeled as a junk topic. In practice, these labels most often reflected what we have termed a domain of communication. For example, the ten emails associated with a topic labeled *finance* might all involve other department managers emailing back and forth with the Finance Department manager to ask them about the proper procedures for purchasing supplies. Alternatively, they might involve the finance department manager emailing with other department managers to clarify budget items and their expectations for their funding needs in the coming fiscal year. A small proportion of topics were associated with emails about some particular event that did not specifically relate to an organizational function (e.g., like managers emailing each other to see if they were all safe after Hurricane Sandy).

The most frequently observed gender mixing pattern exhibits gender heterophily (FM > MF > FF > MM), and does not exhibit a systematic pattern in the topics associated with it. The second most frequently observed gender mixing pattern is male-centric (MM > FM > MF > FF), and is associated with a higher proportion of topics related to planning and public works projects than the other gender mixing patterns. Perhaps most interestingly, the two female-centric gender mixing patterns (FF > FM > MF > MM and FF > MF > FM > MM) are both associated with a large proportion of topics related to finance and HR. As noted above, recent findings by Dasgupta et al. [7] suggest we should expect more female-centric interaction patterns in domains represented by female majorities, and the finance and HR departments are both primarily managed by women. The consistency of our findings with those of Dasgupta et al. are robust across our department and content-based analyses of domain specific gendered communication.

Conclusion

Interactions between employees of an organization represent the building blocks of fundamental processes that drive employee and organizational performance, such as leadership, collaboration, and competition. Communication is a ubiquitous form of intra-organizational interaction. In the current paper we study electronic communication within local government organizations, with the goal of identifying domain specific patterns of gendered interaction. Using internal email corpora covering department managers in seventeen North Carolina counties, we find that the specific forms of gender dependence vary widely across domains. Using two distinct analytical approaches, we find robust patterns in terms of how interaction depends upon gender across domains. Specifically, in communications relevant to finance and human resources, we find female-centric interaction patterns, in which both males and females are more likely to direct their communications to females than to males. Given that the finance and HR department managers in our sample are overwhelmingly female, our results agree with recent findings in the literature indicating that females are more active in small groups when the groups are majority female.

ACKNOWLEDGMENTS. This work was supported by US National Science Foundation Grant CISE-1320219 (Hanna Wallach and Bruce A. Desmarais, PIs). Matthew Denny was also supported under National Science Foundation IGERT Grant DGE-1144860, "Big Data Social Science".

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. Nilanjana Dasgupta, Melissa McManus Scircle, and Matthew Hunsinger. Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering. *Proceedings of the National Academy of Sciences*, page 201422822, 2015.
8. 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.
9. Colin Duncan and Wendy Loretto. Never the Right Age? Gender and Age-Based Discrimination in Employment. *Gender, Work & Organization*, 11(1):95–115, 2004.
10. 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.
11. 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.
12. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
13. 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.
14. Herminia Ibarra. Homophily and Differential Returns : Sex Differences in Network Structure and Access in an Advertising Firm. *Administrative Sceince Quarterly*, 37:422–447, 1992.
15. Peter Krafft, Juston Moore, Bruce A Desmarais, and Hanna Wallach. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems Twenty-Five*, 2012.
16. Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.
17. Belle Rose Ragins and Eric Sundstrom. Gender and Power in Organizations: A Longitudinal Perspective. *Psychological Bulletin*, 105(1):51–88, 1989.

Table 7. Hand coded topic labels, and Topic top-words (one topic per line) for the three topics with the greatest N_t in each cluster, for each cluster associated with each gender mixing parameter ordering.

Coding	FM > MF > FF > MM	Coding	MM > FM > MF > FF
Emergency	fire, cell, drawer, contract, marshal, fax	**Finance**	finance, director, fax, phone, ext, cpa, letter
Elections	public, board, director, email, address, box	**Finance**	insurance, renewal, liability, fax, director
Public Works	water, department, director, meter, kill	**Soc. Serv.**	good, services, director, exercise, department
Health	class, benefits, plan, insurance, enrollment	**Manager**	office, work, time, meeting, today, monday
HR	safety, management, understanding, training, law	**HR/Manager**	leave, work, week, time, pay, years, good
Junk	office, good, airport, call, meeting, fax, time	**IT**	electronic, mail, intended, error, received, recipient
IT	junk, box, summary, emails, report, personal, visit	**Manager**	office, email, time, staff, meeting, work, good
Public Works	director, communications, emergency, central	**IT**	electronic, mail, intended, email, message, recipient
Outreach	health, navigation, attached, main, read	**Health**	health, department, project, email, code, garden
Transit	email, received, intended, confidential, error	**Junk**	email, intended, confidential, error, received
IT	law, box, enforcement,email, records, subject	**Comments**	jail, mobile, inmates, ago, money, jails
Finance	office, finance, wrote, oct, tue, meeting, nov	**Elections**	meeting,email, time, wrote, dec, good, letter
HR	employees, time, on-call, tax, exempt, wrote, work	**Library**	library, fort, time, book, story, thursday
Budget	budget, year, meeting, board, employee, cost, july	**Library**	books, april, free, friends, songs, sale, saturday
Tax	ordinance, changes, manager, email, send, additional	**Planning**	east, planning, street, court, administrator, ext
Planning	meeting, economic, development, planning, office	**Tourism**	fort, year, director, street, full, main, holiday
Emergency	message,email, intended, attachments	**Animal**	outreach, call, animals, inspection, works, public
Manager	main, street, east, fax, manager, office, meeting	**HR**	time, work, issue, full, going, position, hours, budget
		Public Works	public, nashville, suite, washington
		Public Works	email, energy, carolina, north, address
		Public Works	public, nashville, chiller, washington
		Planning	description, director, street, church, suite
		Planning	description, fax, phone, director, street, church
		Zoning	board, meeting, planning, amendment, commissioners
		Emergency	operations, emergency, director, lines, street
		Development	description, director, development, projects
		Tax	office, attached, bill, amount, year, motor, program

Coding	FF > FM > MF > MM	Coding	MF > FM > FF > MM
Finance	order, time, good, april, attached, requests	**HR**	class, benefits, plan, insurance, enrollment, benefit
Finance	budget, phone, finance, media, ext, department	**Planning**	planning, department, box, phone, planner/section
Health	meeting, going, fyi, tricaster, health, project	**HR**	safety, understanding, management, law, enforcement
Finance	meeting, box, fax, finance, attached, resolution	**Spam**	computer, excellent, work, opportunity, applicants
Finance	equity, fax, debt, refunding, time, finance, call	**Inspections**	code, director, office, enforcement
Finance	debt, box, fax, finance, policies, contract, audit	**Extension**	extension, good, road, program, wrote, suite
Finance	learn, leader, director, washington, dream	**Finance**	requisition, approval, munis, department, pending
Finance	fax, ext, phone, finance, director, street	**Manager**	manager, meeting, employees, project, plan, impact
Finance	washington, street, finance, actions, inspire, ext	**Finance**	contract, copy, grant, additional, finance
Health	public, health, email, department, contact		
Health	public, health, email, contact, disclosure		
Finance	good, time, increase, call, pay, office, today		
Manager	fax, east, street, office, main, manager, fyi		
Budget	manager, street, main, fax, office, east, budget		
Budget	fund, budget, balance, year, funds, pay, original		

Coding	FF > MF > FM > MM	Coding	MM > FF > MF > FM
Finance/HR	dss, office, time, cost, allocation, finance	**HR**	class, benefits, plan, insurance, enrollment, benefit
Finance	budget, park, year, salaries, water, bethlehem	**HR**	planning, director, department, meeting, fax, box
IT	password, change, reminder, expires, account, days	**HR**	safety, management, understanding, training, law
Emergency	director, phone, report, fax, emergency	**Soc. Serv.**	intended, message, services,email
Finance	transportation, public, director, legion, read	**Extension**	area, extension, cooperative, street, court
HR	read, year, fiscal, worker, audit, comp, increase	**Finance**	style, questions, span, revenue, fund, week
Planning	public, email, manager, box, board, attorney, law	**Finance**	recipient, attachments, services, requested, email
HR	director, box, planning, phone, fax, resources, human	**Emergency**	office, cell, email, fax, services, emergency
Hurricane	monday, fund, hurricane, island, water, winds	**Finance/HR**	time, call, morning, budget, changes, payroll
Finance	public, mail, electronic, message, time, review	**Health**	health, phone, subject, public, address, third
HR	director, letter, bill, copy, years, june, position	**Health**	health, phone, director, public, disclosed
Inspections	library, property, people, time, list	**Budget**	tuesday, july, animal, commissioners, money
Utilities	public, utilities, facilities, director	**Health**	health, department, street, tel, mph, college
Finance	water, loss, unaccounted, report, gallons, amount	**Finance**	pending, finance, letter, case, director, office
Finance	balance, finance, fund, box, officer, read	**Finance**	requisition, approval, munis, department, pending
HR	policy, office, emergency, director, services, time		
Finance	center, rural, recreation, grant, director		
Tax	tax, administrator, library, year, extension, budget		