Abstract Outline

MATTHEW DENNY SATURDAY 19TH SEPTEMBER, 2015

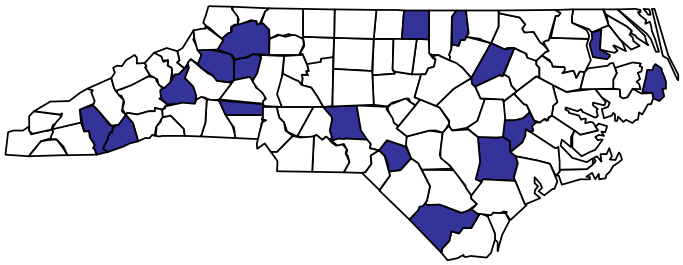
1 Gender in Organizations

1. Gender bias in organizations is well documented in terms of pay, prestige, position, and social interaction.
2. Gender equity is normatively important.
3. Scholars have sought to understand the roots, extent, and nature of gender bias, but have not had access to primary source data. This has traditionally been observational, ethnographic and self reports, which can be biased. Limited in scope.
4. With the increasing use of electronic communication, and the rise of e-government/ transparency, we are now able to use government email to study gender bias using primary source data.
5. Therefore, we choose to explore the relationship between gender and communication patterns in government organizations. This also makes our analyses replicable.
6. To do this, we collect and analyze large scale email data from a sample of 17 local governments.
7. What do we look for, how do we do it, what do we find – at a high level.

2 Data

1. North Carolina has robust public records laws which let us collect data.
2. 22 complied with our request but only 17 counties did so in a way that was useful to our research, allowing us to compare across organizations.
3. Provide some descriptive statistics of the data.

Figure 1: North Carolina county map.



	Manager Gender		Email Sender	
County	**Male**	**Female**	**Manager**	**All**
Alexander	12	9	907	11,924
Caldwell	12	8	121	
Chowan	12	11	2,027	11,737
Columbus	14	10	920	12,707
Dare	15	12	2,247	
Duplin	13	14	1,914	
Hoke	13	11	1,106	5,565
Jackson	18	6	1,499	
Lenoir	15	5	560	10,499
Lincoln	15	7	573	8,727
McDowell	12	5	326	3,494
Montgomery	8	10	680	2,465
Nash	11	8	1,147	9,133
Person	12	9	1,491	14023
Transylvania	16	4	1,857	14,088
Vance	10	8	185	4,349
Wilkes	15	2	303	8,443
Totals:	362	139	17,863	117,154

Table 1: Participating county email statistics. Note that in this study, we only make use of the internal (manager to manager) email data. Some email **All**'s are omitted due to challenges in determining which emails (not sent by managers) were valid in these counties.

4. These counties are a representative sample and we have tons of data so lets start analyzing it.

3 Descriptive Analysis

1. We want to see if there is a relationship between sender gender and recipient gender.
2. If we want to know about gendered patterns of communication in organizations, we should start by looking at aggregate statistics by gender:

	Male	**Female**
Proportion of Managers in Sample	61.6%	38.4%
Average # Emails Sent	48.3	51
Average # Recipients Per Email Sent	1.45	1.43
Average # Emails Received	70.8	71.6

Table 2: Manager email statistics by gender.

3. To test for independence, we construct a contingency table of email sender gender against email recipient gender. We then perform a χ^2 test for independence between the rows and columns.

	Male Recipient	**Female Recipient**	**Total**
Male Sender	7,299	6,286	13,585
Female Sender	5,325	3,510	8,835
Total	12,624	9,796	22,420

Table 3: Manager gender contingency table.

4. The test statistic is $\chi^2 = 92.9$ with a p-value < 0.00001, indicating that the gender of an email sender and its recipients is not independent.

5. We have replicated previous research, we do see gender bias, we could stop here but because we have more information we are going to dig deeper.

6. One place the literature tells us to look for gender bias is in the gender representation in different roles/positions in the organization.

	Emergency	Health	IT	Manager	HR	Library	Plan/Dev	Deeds	Parks/Rec	Finance	Soc_Serv	Veterans	Util/Waste	Elections	Sheriff	Info	Tax	Inspections	Animal	Maintenance	Seniors	Transport	Environment	Misc	Extension
Female	2	11	2	2	12	8	3	9	5	12	8	7	1	11	1	5	5	3	3	0	6	1	4	1	5
Male	15	5	11	15	3	3	11	6	9	5	8	5	11	2	16	2	10	11	9	5	2	6	7	5	8

Table 4: Gender composition of department managers aggregated across counties

7. We can see that some departments are mostly men – the county manager (the boss), the sheriff, and the emergency manager, for example. Others are mostly women – the HR, finance, health, for example.

8. So there is some evidence of gendering in positions, but what about men and women who perform the same position?

9. What we want to know is whether the gender of an email sender and its recipients is independent of the their respective departments.

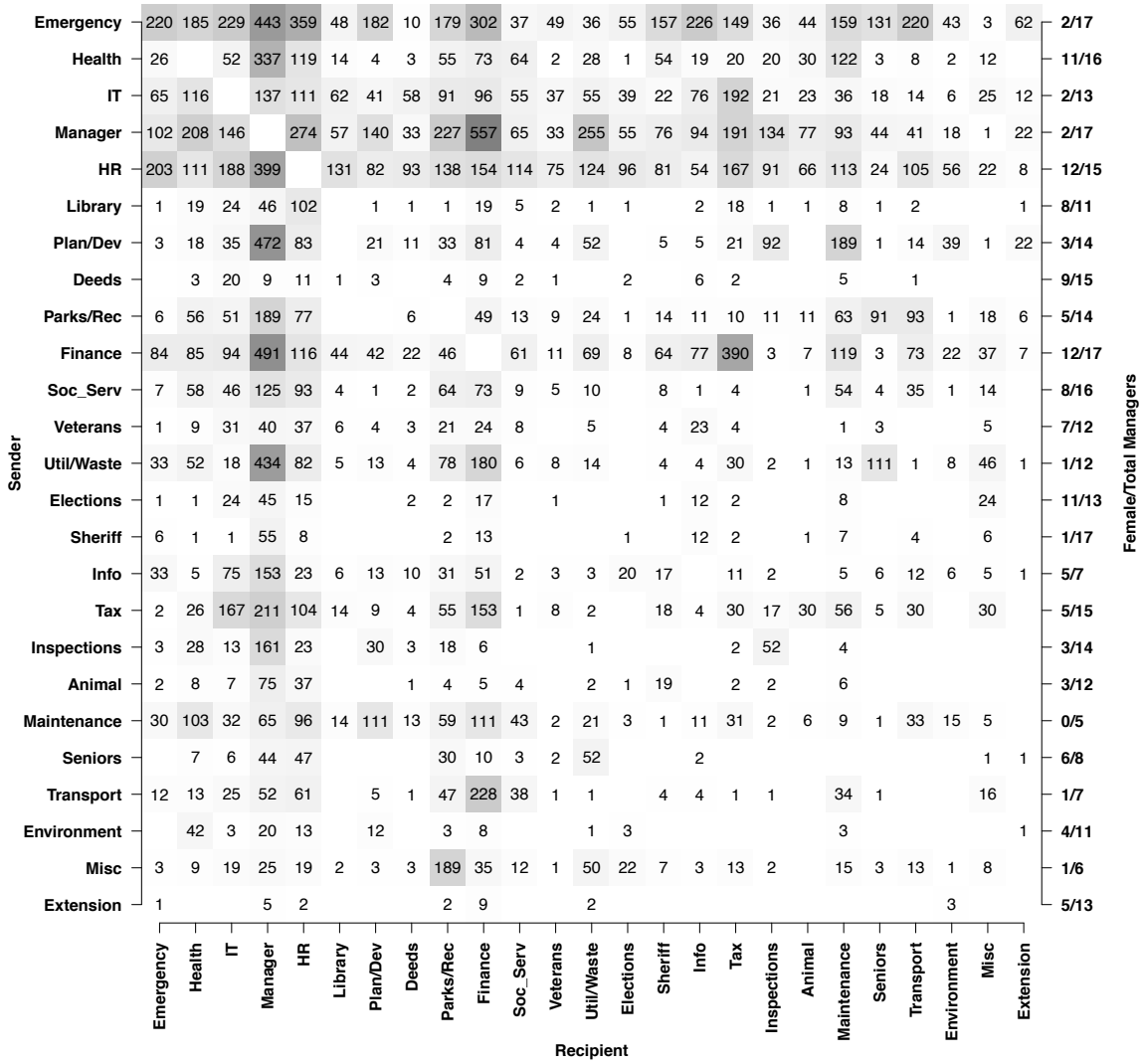


Figure 2: Heat map depicting the number of emails sent from the row department to the column department aggregated across counties. Departments were hand coded into one of 25 different categories based on given titles, to group departments that perform a similar function. The right margin displays the number of counties that had a manager of type X, and the number of those managers who were women.

10. To test for independence, we construct a contingency table of department dyad types (eg. finance – HR) against gender dyad types (eg. male - female). We then perform a χ^2 test for independence between the rows and columns.

11. The test statistic is $\chi^2 = 37,404$ with a simulated p-value[1] < 0.00002, indicating that the gender of an email sender and its recipients is not independent of the their respective departments.

12. This indicates that there is gender bias in communication when we disaggregate to the department dyad level, however, we are still stuck.

13. This analysis does not allow us to disentangle the two potential sources of the pattern we find – bias in the positions women hold within these organizations, and bias in the way that male and female managers communicate.

14. Furthermore, this analysis may be slicing our data too thin since not all counties have a department manager of every type, and any particular department dyad may only exchange a handful of emails. Moreover, the department attribute of each manager is likely caused by several factors other than the topics of communication that may relate to gender bias.

15. One solution to the problems raised above is to model the email content, as general topics such as balancing budgets are commonly represented across counties. Practically, this also allows us to focus on a smaller number of communication topics that are shared across counties.

[1]We use bootstrapped p values calculated using 50,000 resamples as they are more conservative.

16. To do this, we might want to use a statistical topic model to categorize emails, then model gender mixing using the LSM in each topic.
17. However, the gender mixing parameters we infer using this approach would be confounded by selection effects. As a toy example, we would not be able to tell whether men prefer to talk to male coworkers over female coworkers about football, or women prefer not to talk about football at work, and so they do not participate in those conversations.
18. To disentangle these effects, we need a joint model for email topical content, and the structure of communication.

4 A Model of Email Content

1. Motivation for building a model for email data – who you send an email to depends on what it is about. And what you say depends on who you are talking to.
2. Our solution: a generative model for email topics and recipients. Then discuss the existing TPME model and why we build on it.
3. Overview of the generative process in plain english.
4. Describe the generative process for LDA part of model.
5. Model based topic clustering.
6. Explain how draws of message recipients are conditioned on the email topics.
7. Describe the generative process for the latent space portion of model.
8. Summarize the generative process and lead into inference.

4.1 Inference

1. We have to invert the generative process to perform inference on the model parameters.
2. We use block Metropolis Hastings within Gibbs sampling.
3. A beta R package is available for those interested.
4. Discuss our model specification and justify our hyper-parameter choices.

5 Gender Mixing

1. Since we have modeled content and structure together, we have broken up the confounding we were concerned about, and we can now interpret the gender mixing parameters that come out of our model with more confidence.
2. Lets return to the question we posed at the end of our descriptive analysis section: is there gender bias in the patterns of communication in these county governments.
3. Add in Bruce's MV KS test results for each county here.
4. Interpret the results.
5. Take as a case study, a county with the highest degree of gender bias, as identified by this method.
6. interpret the topic model output from two different clusters and discuss.