Who To Reply To First? (But First: Who Would You Curse To?)

Alec McGail

***Abstract*—The first goal of this project was to assess whether and how the presence of cursing is related to the relationship of the interlocutors. To investigate this I have analyzed two datasets of organizational emails. Each dataset contains emails from a subset of employees at either Enron [3] (500k emails) or Avocado [7] (1.2M emails). It turns out that in context of emailing from a work email address, the presence of cursing can be used to identify strong and informal relationships. Using topic modeling, for example, we can see that the topic of discussion among those who share a curse is far different from the norm.**

The second goal of this paper is to leverage email timestamps and uncover judgements made by users as to which emails to answer first. I justify that these measures hold significance and that they indicate preference relationships, at least in aggregate. These theoretic justifications are further qualified through a semantic analysis of the emails, using LDA to generate a topic classifier for sentences. Some topics correlate with the measures I have presented and seem to discourage or encourage replies in aggregate.

Keywords*—*cursing, intimacy, nothing too great

I. Introduction/Motivation

My initial motivation was to understand whether cursing really does encourage friendship, as is folk wisdom. Although cursing is an amazing phenomenon within NLP – easy to detect, socially impactful and full of meaning – cursing shows up infrequently in these datasets, as shown in Figure 1. This prevents me from saying very much about cursers vs. noncursers. I will give a brief analysis on cursing, showing that those users who are comfortable cursing with each other interact differently than the population at large.

The main focus of this paper will be on the construction of a subjective relative value order on emails based on the timestamps of the original message and its reply. This measure addresses several problematic aspects of comparing the time to response by itself. In particular, individuals are not always checking their email. For example, it's quite hard to say that an email taking 6 hours for a reply necessarily means it was neglected, as the person may simply not be answering emails in general for this period. Likewise 1 hour may be a "longer" time if they are spending that hour answering other newer emails.

H1 The presence of cursing in conversation indicates intimacy.

H2 The topics discussed in an email are correlated in aggregate with the priority the recipient places on giving a reply.

H3 Shorter emails get quicker replies, being easier to read. (I've included this because it's intuitive, and quite easy to test.)

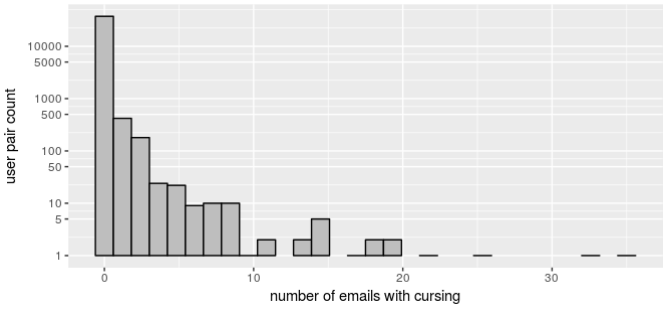

Fig. 1. Histogram showing the number of times each pair of users uses a curse (any of those listed in the Methods section).

II. Data Description

A. The Emails

The Enron and Avocado datasets (hereafter *En* and *Av*) consist of the raw text of all emails which were on the hard-drives of the employees' computers at the time of the data scrapes. Neither claim to be absolutely complete. If all individuals involved in an email delete it permanently from their computer, it will not be retained in the dataset. Furthermore, *En* has gone through a process of redaction upon request, which has removed an undocumented number of emails, and through processing for release, some emails (e.g. from or to lawyers) have been removed altogether. I should also mention that I am required not to present any content whatsoever of *Av*, so I will restrict its use to comparison and network analysis in findings. Any illustrative text, e.g. from a topic model, will come from Enron.

Each email includes the header, which all have at least the tags: *Message-ID*, *Date*, *From*, *To*, *Subject*, *Cc*, and *Bcc*. In *En*, 200 emails have an associated *Date* before 1990 or after 2005, likely due to an incorrect date on their computers. *Av* contains 77 such emails. In both datasets these were excluded. To canonicalize *From* and *To* I used a RegEx built for identifying RFC 5322 compliant emails[1], slightly extended to include some aberrant internal email addresses. For *From* I canonicalize to the last match of this pattern, and for *To* to the set of unique email addresses which match. Note that this may fail to give what we want for a *To* field like

```
am2873@cornell.edu <amcgail2@gmail.com>
```

but fields of this form were so few that I neglected to address the issue.

[1] http://emailregex.com/

En has 520,935 email files and *Av* has 1,415,574. Because each email sent between users within the company may be stored on both interlocutors' computers, there are duplicates which must be removed. I identified duplicates as any emails with the same sender and recipient, and exactly the same time sent. Through this reduction of redundancy, *En* is narrowed to $242,908$ emails and *Av* is narrowed to $538,926$. It's rather shocking to me that in every description or analysis I've seen of either of these datasets this step isn't mentioned, and the entire analysis is carried out with the original sample sizes. This leads me to believe that they used duplicates in their analysis, although I haven't confirmed this.

The next crucial step was to pair replies to the original email they replied to. The method I used for this is the following:

- Create a new field for each email, *subjectStripped*, which contains the *subject* field, modulo a possible prefix of the form "`[R][Ee]:\w*`". There are no instances of multiples such prefixes in either dataset, so more complication isn't necessary.
- Group all emails by *subjectStripped*
- For each group g, for each email e, find another email from the recipient of e to the sender of e which also belongs to g, with timestamp closest to e, but which e precedes.
- Because I don't analyze group conversation in this analysis, I only do this identification for emails with a single address in *To*.

Of course we don't know whether they were replying to e or were in the middle of replying to an earlier email in g while e arrived. They could have also happened to craft an email with exactly the same title as e, before any conversation started (this is plausible for the empty subject). We also should be aware that replies in which the replier changed the subject will not be identified.

This method identifies $18,758$ replies in *En* and $59,636$ in *Av*, approximately a 10% sample. Although this might appear unreasonable at first sight, it becomes more believable with the presence of spam, and of emails which don't need a reply, in addition to the fact that the vast majority of emails which receive a reply receive exactly one.

For later analysis I also needed to extract the content in the body of the email which the user wrote themselves. A reply typically contains the full body of the email it's replying to, completely confounding any meaningful linguistic analysis. I also removed excessive automatic footers, and any large body of text which is repeated without user input. The process to remove such content was trial and error, writing a RegEx for each new class of separators which divide the true content the user typed from everything else. The production set I used was the following:

```
.*Original Message.*
[^\n\r]\s*-+.*wrote:$
[^\n\r]>>>.*
[^\n\r]---.*
[^\n\r]\s*>?(From|To):
<html>
<!--
\*{10}
```

I then selected 100 documents randomly and checked whether anything had been cut off which shouldn't have, or there was unwanted text which should have, and found that this production set processed the emails near-perfectly. Some of the RegEx's would leave a few words behind in a small number of cases, but the extra effort to get these is not worth its impact on the analysis.

B. Power Hierarchy

For some analyses I will pair *En* with Agarwal et al.'s [1] determination of hierarchy within Enron. This pairing has extreme limitations, mostly that it is not complete among those who email most in *En*. It contains a mapping from person to email address for 1518 Enron employees (most individuals have many email addresses, which change throughout their tenure), and from person to position within the company. It also provides a directed hierarchy with edges "manages" and "supervises" between positions. I was able to extract 992 immediate occupational relationships, with 585 interpersonal relationships mapping onto these. As in [1] I compute the transitive closure of this graph, bringing the total to 786. These numbers are much lower than the $2,155$ and $13,724$, respectively, mentioned in Agarwal et al., indicating I've probably done something wrong. By joining these 786 dominance relationships with the various email addresses for each individual I got $10,066$ dominance relationships between email addresses. All in all, this maps onto only $4,602$ emails between two individuals out of the $179,016$ emails of this kind.

III. Methods

A. Cursing

To identify cursing in the datasets, I restrict myself to those most heinous curses (and thus the most socially stigmatized). I also refrain from using words which commonly occur as part of other words or in contexts in which they are not curses. The list I settled on is *fuck shit douche cunt pussy jesus christ dumbass asshole tits damn slut* and *dammit*.

Although 46% of emails in *En* are between only two people who both have "enron" in their email address, only 26% of emails between only two people with the word "fuck" in their body have this property, our first indication that a work stigma exists against cursing, but I think a surprisingly weak one. Only two of these emails matched to a dominance relationship. For *Av* the corresponding percentages are 57% and 34%.

To understand the nature of these emails, and how they differ from the general body of emails, I've generated a topic model over the two groups of emails. These are shown in Table II and I. Identifying

B. Reply time

Are there lexical attributes of an email which correspond to shorter or longer response times? To answer this question I will run through a few different methods, attempting to refine the metric of "response time". A histogram of these response times is shown in Figure 2. The first such method is to simply

Topic	Example words
Topic 0	follow sent email week message regards
Topic 1	person look forward want good process
Topic 2	mobilize plans applications capabilities
Topic 3	solution resources running enables short
Topic 4	million round solutions technology ventures
Topic 5	sure make services market need aware
Topic 6	product mail address called person application
Topic 7	speak help like regarding working opportunity
Topic 8	working group currently business initiative
Topic 9	think work good right today week office
Topic 10	meeting like meet schedule week technology
Topic 11	send phone want number email info enabled
Topic 12	recently available year package completed
Topic 13	sprint opportunity santa clara include
Topic 14	internet stock list watch report people
Topic 15	trade airlines future provide companies
Topic 16	including content enable provider post
Topic 17	kelsey touch plans follow asked pricing
Topic 18	bank online company said soon interactive
Topic 19	thought think article really folks area
Topic 20	contact questions free feel happy provide
Topic 21	access services going telephone site
Topic 22	company partners consulting microsystems
Topic 23	lead account channelwave leads sales partner
Topic 24	applications business access online customers

TABLE I. A TOPIC MODEL OVER THOSE EMAILS IN *En* WHICH DON'T CONTAIN THE WORD "FUCK"

Topic	Example words
Topic 0	need trying send work friend does getting mail chill
Topic 1	f*cking want guys taken thinking sales care think apartment
Topic 2	mail great things monday want girls prices auctions school
Topic 3	f*cking email sh*t didn going went sorry f*ck people
Topic 4	f*ck talk said thought meeting tomorrow thing told cool
Topic 5	sure actually f*cks good people horse make feel said
Topic 6	avocadoit tell times check going usually does yesterday good
Topic 7	place need night minutes cause thing rocked today server
Topic 8	doing working home look time make office having feel
Topic 9	free year sh*t getting girl sent mean include company
Topic 10	email intended recipient mail message doing confidential
Topic 11	time think pretty phone internet seen long crazy gonna
Topic 12	want money didn going guess phil hook later bunch
Topic 13	week f*cking live coming hard click f*ckin difference read
Topic 14	little weekend really year think work come long going

TABLE II. A TOPIC MODEL OVER EMAILS IN *En* WHICH CONTAIN THE WORD "FUCK".

take the actual difference in timestamps between the original message and its reply as a measure of precedence. I then split these replies into two piles: 1) those replies which occur less than ten minutes after the original, and 2) those replies which take between 10 and 15 hours. I also limit each original emailer to five emails per pile, to reduce the chance that a single person is overrepresented in the model, and each email to the first five sentences. I also semantically parse each sentence to weed out non-sentences, such as a person's email signature. To do this I identify POS of each word in the sentences, and exclude sentences for which the number of words with

$$\text{POS} \in \{\text{NOUN, VERB, PROPN, DET}\} \cup \{\text{PART, PRON, ADP, CCONJ, ADV}\}$$

divided by the number of other "abnormal" words is less than 1.6, a number again determined through trial and error. This leaves me with $2,664$ sentences in the short response pile, and $2,304$ sentences in the long response pile. For comparing

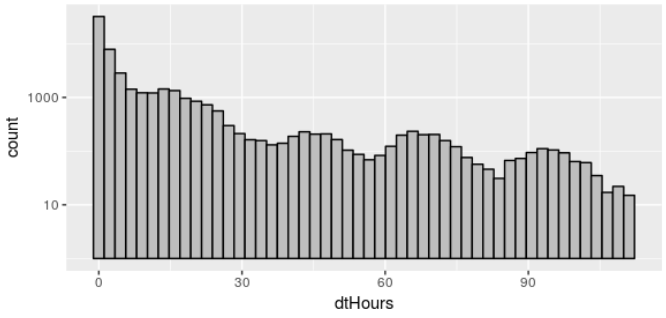

Fig. 2. Histogram of response times in *Av* in hours. The y-axis (count) is logarithmic.

two lexicons, I use a simple model of content creation. That is of one group pulling their vocabulary from a uninformative Dirchelet prior over multinomials with parameter $\alpha = 0.01$, and with the words being drawn from this vocabulary distribution. For more information see Monroe et al. [6]. From here forward I will refer to this analysis as "Fightin' Words," in reference to the paper. In this first analysis, I exclude words which show up in less than 1% of sentences, focusing on relatively common words. The results of this analysis are shown in Table III.

In the second analysis, I exclude words which show up in less than 0.1% of sentences, or more than 20% of sentences, focusing on more specific words. See Table IV.

I feel that this analysis wasn't able to show specific topical trends, but only large scale focus trends. When I allowed lower-frequency words in, the results at face-value seem rather spurious. I will address this by constructing a topic model over emails in *En* and comparing the topics of emails to the reply *precedence*, a metric I define in the next section.

	word	zscore
1	of	-3.04
2	review	-2.66
3	an	-2.55
4	attached	-2.48
5	and	-2.38
6	tomorrow	-2.32
7	week	-2.25
8	with	-2.22
9	more	-2.22
10	him	-2.17
11	hi	2.05
12	at	2.17
13	information	2.26
14	your	2.37
15	if	2.64
16	can	2.97
17	you	4.54

TABLE III. A LARGER *zscore* INDICATES THE WORD IS MORE PREVALENT IN THE SHORT RESPONSES PILE. THESE RESULTS SEEM VERY INTUITIVE TO ME. FOR EXAMPLE "HI", "INFORMATION", "YOU" AND "YOUR" ALL INDICATE SHORT RESPONSE TIMES, WHILE "REVIEW", "ATTACHED", "TOMORROW" AND "MORE" INDICATE LONG RESPONSE TIMES.

	word	zscore
1	of	-3.43
2	review	-2.92
3	capacity	-2.77
4	version	-2.74
5	week	-2.66
6	operations	-2.60
7	assembly	-2.54
8	proposal	-2.53
73	michelle	2.56
74	deals	2.57
75	information	2.60
76	october	2.63
77	meter	2.63
78	request	2.63
79	stay	2.66
80	your	2.70
81	reply	2.71
82	mw	2.81
83	month	2.82
84	joe	2.85
85	if	2.85
86	can	2.93
87	yesterday	3.06
88	feb	3.09
89	thank	3.25
90	east	3.29
91	order	3.65

TABLE IV. FIGHTING WORDS ON MORE SPECIFIC (LESS FREQUENT) WORDS. SEE THE CAPTION FOR TABLE III FOR MORE DETAILS

.

	Phrase	z-score
0	know your thoughts	6.38
1	to look at	5.37
2	<employee name> sent	5.15
3	trying to find	4.41
4	want to clarify	4.27
5	if we can	3.97
6	you would prefer	3.77
7	the conference call	3.75
8	hope you had	3.54
9	talking points for	3.45
10	to pay the	3.23
11	member of tgf	3.20
12	conflicts with the	2.91
13	would not be	2.39
14	proceed with the	2.19
15	has asked me	-2.98
16	<employee name>	-2.58
17	detailed activity list	-2.58
18	new buyer 14jan02	-2.58
19	in relation to	-2.55
20	on the analyst	-2.46
21	to it yet	-2.46
22	say thanks for	-2.19
23	delete this message	-2.10
24	something we can	-2.10
25	to speaking with	-2.06
26	in london and	-2.03
27	send you copy	-2.03
28	<pager number> pager	-2.03
29	is as follows	-2.02

TABLE V. "FIGHTIN' WORDS" ANALYSIS ON TRIGRAMS. A RANDOM SAMPLE OF 15 POSITIVE SIGNIFICANT Z-SCORES AND 15 NEGATIVE Z-SCORES. A POSITIVE Z-SCORE INDICATES ASSOCIATION WITH PRIORITIZED EMAILS, AND A NEGATIVE Z-SCORE INDICATES ASSOCIATION WITH DEFERRED EMAILS.

C. Reply precedence

I would like to differentiate whether a long response time is due to an explicit preference, or because the replier was simply not answering emails in that time (e.g. over a weekend, or during a more-than-busy day). In this spirit, I will look at the focal emailer's *relative* response time to other emails in their inbox. That is, a question of "how long did she take to answer the email in question" turns into "how many other emails did she prioritize over the email in question".

I define a focal emailer P to have **prioritized** e^+ over e^- (I'll say $e^+ >_P e^-$) if P received e^+ after e^-, but P responded to e^+ before responding to e^-. This means that for the whole time e^+ was in P's inbox, e^- was too. I do not define this relationship for any emails which don't receive replies. I say P has **deferred** an email e if they have prioritized another email over it. The concept *prioritized* pairs emails very naturally for analysis, where a single user replied to two emails in the same general time period, but seemed to prioritize one email over the other. There's no question that any specific assigned priority of this sort could be spurious, that it was simply because they saw the notification pop up, and had forgotten about the other. However, when linguistic patterns emerge from this relation, or when an individual *consistently* "prioritizes" or "defers" certain emails, as defined above, it's very likely that these concepts coincide with their homonyms.

In *En* there are $5,348$ such priority pairs among 258 focal emailers, with a maximum of 620 such pairs for a single focal emailer. In *Av* there are $36,859$ such priority pairs among 299 focal emailers, with a maximum of $4,553$ such pairs for a single focal emailer.

I've run the same "Fightin' Words" analysis, except on word triples. This was in the hopes that it would reveal some more contextual information, and I think it did. I've presented a random sample of significant words and phrases in Table V

—

D. What topics deserve speedy replies?

To answer this question I first extracted the email pairs (o_i^+, o_i^-) such that $o_i^+ >_{P_i} o_i^-$, where P_i is the recipient of the emails o_i^+ and o_i^-. I fit an LDA model to the collection of sentences extracted from the body of these emails, as a baseline model for what these employees talk about. I exclude all terms which occurred in over 20% of documents, in order to isolate more domain-specific words, or which occurred in under 1% of documents, to eliminate anomalies, and after performing a grid-search on the number of clusters on the set $[1, 150]$, I decided on 15 clusters based on the model's log-likelihood. The resulting topic model over *En* is shown in Table VI.

For each email $e \in \{o_i^+\} \cup \{o_i^-\}$ I used this LDA model to generate a vector of "what they talked about". To do this I took each sentence $s \in e$ and predicted which category it's in. First I projected the sentence onto my LDA model, generating a probability distribution over classifications, p_c. Let $c^* = \text{argmax}_c(p_c)$ be the most probable classification. If $p_{c^*} < 0.7$ I discard the choice of classification. Let $\text{odds}(c^*) = p_{c^*} / \sum_{c \neq c^*} p_c$, the odds of c^* versus any other

classification. Then I defined a confidence score in $[0, 1]$ by

$$\text{conf}(c^*) = \frac{1}{1 + \text{odds}(c^*)^{-1}}$$

This choice was based on multiple needs. I'd like

$$\lim_{s \to 0} \text{conf}(s) = 0$$
$$\lim_{s \to \infty} \text{conf}(s) = 1$$

$\text{conf}(s)$ is continuous and monotonic, and is linear near $s = 0$. I chose $\text{odds}(c^*)$ as the argument, because I'm picking only one classification for this sentence (assuming here that a sentence can only be about one thing), so I'd like c^* to be much more likely than any of the others. Also, if the LDA is entirely uncertain it will return a uniform distribution, in which case $\text{conf}(c^*)$ will be small, as desired.

Now that I have a predicted category and confidence for each sentence in the email, I represent the email by a vector $\mathbf{C}(e)$, where $\mathbf{C}(e)_c$ is the sum of confidences for each sentence classification for category c. To understand the difference in *classification* between emails which were prioritized and emails which were deferred, I compute the sum of these vectors for each category, being careful there are the same number of sentences from each email pair. Define

$$\mathbf{W}^+ = \sum_{o_i^+} \mathbf{C}(o_i^+) \text{ and}$$
$$\mathbf{W}^- = \sum_{o_i^-} \mathbf{C}(o_i^-)$$

You may feel I'm leading you far afield, but I wouldn't be doing so if this intuitively-grounded NLP analysis wasn't leading anywhere. I've listed in Table VII metrics describing the difference in topic distribution between preferred and deferred emails. The most deferred topics are 2 and 10, which seem to be topics representing the request to review an attached document and reply, and to schedule a status meeting, respectively. The most preferred topics are 14 and 11, which seem to be a scheduling / planning topic (but with "forward" and "look" as the top words) and a request for comments and questions. Intuitively these seem reasonable!

E. Do shorter messages get quicker replies?

Message length is distributed approximately as a power law in my distribution. I used paired Wilcoxon tests on the message lengths of the preferred and deferred messages to determine if length was correlated with preference. In fact it is, with the means of these distributions being distinct with $p < 10^{-10}$. The difference in character count for the original emails is $\in [-15, -36]$ (95% confidence interval), and the difference in character count for the *replies themselves* is $\in [-23, -34]$ (95% confidence interval).

Topic #	Representative words
Topic 0	think right want need vince market good request make
Topic 1	attached comments copy review draft received today mark issues
Topic 2	agreement enron attached changes request sara review draft energy
Topic 3	enron jeff houston mark mail work friday vince email
Topic 4	know office louise list power enron issues time john
Topic 5	message send just email mail sent received thought list
Topic 6	know change need does plan like work thought going
Topic 7	just know jeff wanted working thanks group today week
Topic 8	credit power company energy trading following group tana want
Topic 9	deal deals date need contract think following number john
Topic 10	like discuss meeting talk things going today monday louise
Topic 11	time know make sure want just comments wanted questions
Topic 12	thanks help jeff know business number tana sent questions
Topic 13	week good morning friday tomorrow meeting today time going
Topic 14	forward look information going getting need think schedule know

TABLE VI. LATENT DIRICHLET ALLOCATION ON 15 TOPICS OVER ALL RESPONSES WHICH ARE EITHER PRIORITIZED OR DEFERRED.

Topic # (c)	$\mathbf{W}_c^+ - \mathbf{W}_c^-$	$\frac{\mathbf{w}_c^+ - \mathbf{w}_c^-}{\mathbf{w}_c^+ + \mathbf{w}_c^-}$
Topic 0	-25.21	-0.06
Topic 1	-4.18	-0.01
Topic 2	-53.56	-0.17
Topic 3	11.18	0.03
Topic 4	11.72	0.03
Topic 5	13.85	0.04
Topic 6	-14.31	-0.03
Topic 7	5.53	0.02
Topic 8	20.21	0.05
Topic 9	30.56	0.07
Topic 10	-61.14	-0.11
Topic 11	49.82	0.10
Topic 12	0.50	0.00
Topic 13	-3.90	-0.01
Topic 14	39.18	0.12

TABLE VII. THE DIFFERENCE IN TOPIC ALLOCATION BETWEEN THE PAIRED GROUPS $\{o_i^+\}$ AND $\{o_i^-\}$

IV. RELATED WORK

Researchers have been playing around with *En* for quite some time, and more recently have been using *Av* as well. Gilbert attempted to extract a lexicon which signals workplace hierarchy using *En* and the hierarchy dataset cited above [4], very similar to goal in Agarwal et al. [1], which uses social network analysis. Alkhereyf and Rambow [2] categorize emails from *En* into "Business" and "Personal" in both *En* and *Av* based on the social network of email communication, along with lexical attributes. Prabhakaran and Rambow [8] explore the connection between gender and the manifestation of power relations in *En*. McCallum et al. train an LDA model for topic on *En*, based on the specific individuals who are corresponding, claiming great success in identifying not only topic but *relationships* of the individuals [5].

V. CONCLUSIONS

I initially chose these datasets because of a single characteristic: the communication was assumed to be private at the time by the individuals involved. This property seemed to me to be ideal for capturing the linguistics of real relationships. In contrast, public data-sources (e.g. communicating on web

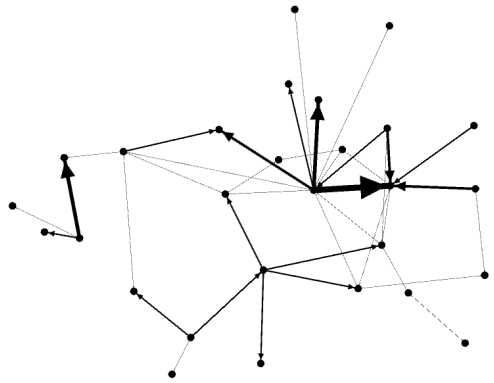

Fig. 3. The focal preference network for an employee at Avocado. Each node represents another employee, and the strength and direction of ties indicates how many emails by one user were preferred over another.

sites) always have an audience which is known and consciously addressed by the data-creator.

This *is* a great fact about *En* and *Av*, however this is trumped by the fact that the emails are drenched in hyper-structured social contact. Most emails are about work, and are between individuals who have formal relationships. The ideal context for studying my original research question is likely something like chat datasets.

Admittedly I have no statistical significance attached to my presented metric, and unless you believe in the methodology I've put forward, I feel much more justification would be needed to rely on these measures (possibly an entire paper or set of papers).

VI. Future Work

A. Relationships

It starts to become apparent in Table V that there might be more than just lexicon which determines who is responded to sooner. For example, phrase 2 seems to have creeped through my preprocessing, and simply identifies a specific person. This could either mean this person replies quickly to emails, or others reply quickly to him. Phrase 28 shows the same thing, but in the opposite direction. Thinking about what relationships are present in this organization can help us in identifying the processes going on, and the previous analysis of prioritized and deferred emails gives a good lense to look at this through.

Considering preferences over emails as preferences over *users* immediately sheds light on this issue. By identifying all these preferences for each focal emailer in the dataset (provided they have enough incoming emails to judge between), we can begin to understand what patterns are arising simply because of the relationships between the users, and not as much because of the language being used.

Figure 3 shows the network derived from such analysis. Many attributes immediately come to the fore. First, there are very large explicit preferences of one user over another. Secondly, there are stars and sinks, nodes for which all (or most) edges are in or out. This indicates strong-positive or strong-negative relationships, and is a good sign for this sort of analysis. Although I wasn't able to incorporate this directly in the current paper, I'd love to analyze it further to get a deeper understanding of what we can and can not retrieve from analysis of emails.

B. Relationships over time

Cursing extracts friendly relationships, and when further restricting to long relationships (many emails back and forth), we typically encounter a progression of relationship, sometimes all the way back to applying for the job in the first place. I think the analysis of these dynamic relationships is extremely interesting, and these datasets offer a great medium for their analysis. Unfortunately it's not something I got to in this analysis, but I'd love to dive in to these phenomena in particular, and characterize linguistically what changes over the course of these relationships.

References

[1] Apoorv Agarwal, Adinoyi Omuya, Aaron Harnly, and Owen Rambow. A Comprehensive Gold Standard for the Enron Organizational Hierarchy. *Proceeding ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, 2:161–165, 2012.

[2] Sakhar Alkhereyf and Owen Rambow. Work Hard, Play Hard: Email Classification on the Avocado and Enron Corpora. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, ACL 2017*, pages 57–65, 2017.

[3] Dr. Michael W. Berry, Murray Browne, and Ben Signer. *2001 Topic Annotated Enron Email Data Set (LDC2007T22)*. Linguistic Data Consortium, Philadelphia, 2007.

[4] Eric Gilbert. Phrases That Signal Workplace Hierarchy. *Proceeding CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1037–1046, 2012.

[5] Andrew McCallum, Xuerui Wang, and Andres Corrada-Emmanuel. Topic and role discovery in social networks. *Journal of Artificial Intelligence Research*, 30:249–272, 2007. ISSN 10450823.

[6] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin ' Words : Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16:372–403, 2008.

[7] Douglas Oard, David Kirsch, and Segey Golitsynskiy. Avocado research email collection. *Philadelphia: Linguistic Data Consortium*, 2015.

[8] Vinodkumar Prabhakaran and Owen Rambow. Dialog Structure Through the Lens of Gender, Gender Environment, and Power. *Journal for Dialogue & Discourse*, 8 (2):21–55, 2017.