1 Introduction

Enron was an energy company based in Houston that misled the public with fraudulent accounting. The scandal was publicized in October of 2001, after which Enron declared bankruptcy.[1] In the course of the ensuing legal investigation, the email folders of several of Enron's senior managers were released to the public, and those emails comprise our dataset. The dataset contains:

- # of users: 150
- # of email addresses: 20,328 email addresses in the `from` field
- # of email addresses from Enron: 6,460
- # of users in both `from` and `to`: 131

The data lends itself to two primary datasets: one which includes *all* emails from Enron users, and one which includes only the emails sent between users whose full inboxes are in the corpus. The former allows for conclusions on overall email activity at Enron, including, for example, how many total emails were received by these 131 users over time. The latter, more limited dataset is required for network analysis, since only in this dataset do we have a complete picture of the users' connectedness. I will refer the the former dataset as the "full" dataset, and the latter as the "network" dataset.

These initial counts introduce several questions worth exploring: How much is one user connected with the others? Do most emails come from just a handful of senders? Do people who send lots of emails send to a larger audience? Which days of the week do people send the most emails?

2 Data

2.1 Initial cleaning

Several email addresses were initially distorted since they were in a slightly different format. I used regular expressions to correct these addresses; for example, I altered `credit <.williams@` (which, in raw emails appeared as a substring of `credit <.williams@enron.com>`) to what should be the correct mail name `.williams`.

2.2 Unique keys

Each user has multiple folders of emails. Information from all of those folders may be interesting, but I had to be careful not to load duplicates of the same email if, for example, an email were in both someone's "inbox" and "favorites" folders. In order to remove duplicates, I captured the email dates, listed in the second line of each email after "Date: ". I loaded these dates using syntax similar to the base code, and named the dates `email_date`. I reasoned that uniqueness on the sender, recipient, and timestamp should uniquely identify emails. I appended a unique ID (a UUID) called `message_id` to avoid using the multi-column key.

[1] According to the Wikipedia article "Enron Scandal" at `https://en.wikipedia.org/wiki/Enron_scandal`.

Importantly, one email sent from one sender to multiple recipients has *separate* message IDs for each recipient. The key `from` and `email_date` uniquely identifies a message that was sent to multiple recipients. Naturally, `email_date` also allows me to plot messages sent over time.

2.3 Reading particular lines

Regular expressions make reading in the subject line of an email fairly straightforward. I removed duplicated subjects to avoid over-representing the subjects of email chains that lasted a long time or were sent to multiple people. A downside of removing duplicates is that I have also removed separate instances of the same subject. Without parsing the emails to identify reply chains, finding unique subject lines is challenging.

3 Results

3.1 Senders & Recipients

The most prolific senders and the biggest email recipients offer insight into the scale of the number of emails sent. Their positions reveal who top communicators were, and could hint at which groups of users are most closely related.

Top senders include Vince Kaminski (Managing Director for Research), the Enron Announcements and 40enron accounts, and Steven Kean (Chief of Staff). It appears that the top recipient `j.kaminski` may be another email address for Vince Kaminski, which should make him the most active email communicator in the full dataset.

Full Dataset

	Top Senders		Top Recipients
pete.davis	3811	j.kaminski	7332
vince.kaminski	3765	.taylor	6031
enron.announcements	2636	jeff.dasovich	5625
40enron	2318	tana.jones	5606
mark.taylor	2042	j..kean	4972
richard.sanders	1604	sara.shackleton	4631
steven.kean	1408	louise.kitchen	3746
announcements.enron	1181	j..farmer	3572
john.lavorato	1122	gerald.nemec	3271
jeffrey.shankman	1013	sally.beck	3127

Network Dataset

	Top Senders
jeff.dasovich	652
sara.shackleton	370
louise.kitchen	367
richard.shapiro	338
stephanie.panus	333
mike.grigsby	330
tana.jones	281
d..steffes	251
rod.hayslett	238
kimberly.watson	199

	Top Recipients
j..kean	493
.taylor	424
tana.jones	401
sara.shackleton	376
michelle.lokay	353
louise.kitchen	267
gerald.nemec	264
kimberly.watson	257
jeff.dasovich	253
richard.shapiro	206

In the network dataset, the counts are unsuprisingly lower, and some senders have been removed. For example, `enron.announcements` did not receive emails, so it is not part of the network dataset. Sara Shackleton (a vice president) is an active emailer in the network set.

3.1.1 Distribution of Senders & Recipients

The distribution of emails sent and received by each user shows that these top senders and recipients were rather far from typical.

Figure 1: Histograms of emails sent and received on full and network datasets

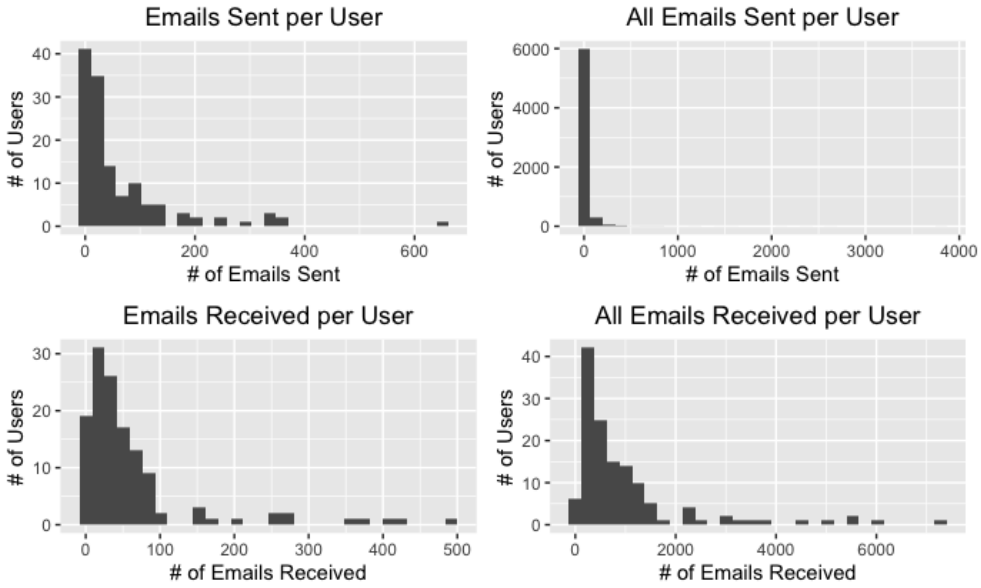


Figure 1 shows the network dataset on the left and the full dataset on the right. Most senders send very few emails, while only a few send many emails. This is particularly true in the full

dataset, since for most senders we have only a handful of emails which have been sent to the 131 users whose inboxes we have. In other words, the picture for the upper right plot is not complete. In contrast with the "emails sent" distributions, the distribution of emails received is a bit wider.

3.1.2 Matrix of Senders & Recipients

The connections in the network set can be visualized to a limited degree via a matrix, labeled by "emails from" vs. "emails to" and colored by the number of emails matching the cell.

Figure 2: Weighted adjacency matrix of emails sent and received

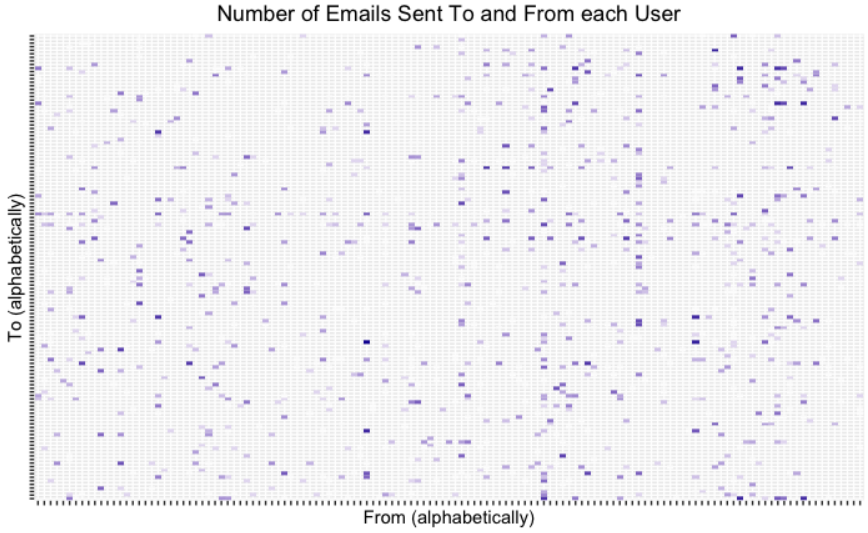


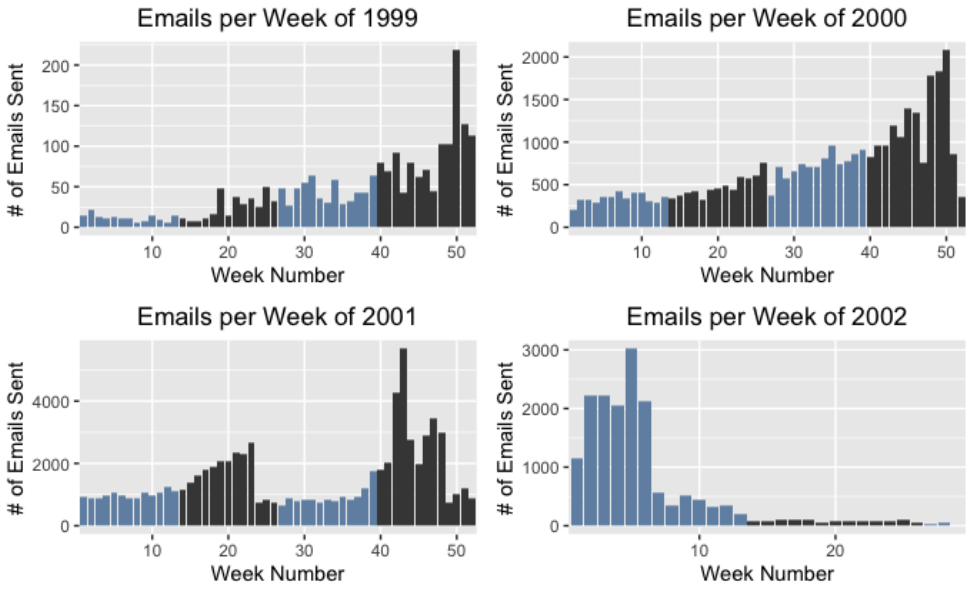
Figure 2 clearly indicates that communication was relatively sparse. The names of users are listed alphabetically but are not labeled due to space constraints. If the matrix were smaller, this could be a more effective visualization.

3.2 Emails Sent Over Time

3.2.1 Enron Timeline

The number of emails sent per week chronicles the fall of the Enron corporation.

Figure 3: Emails sent by week of each year



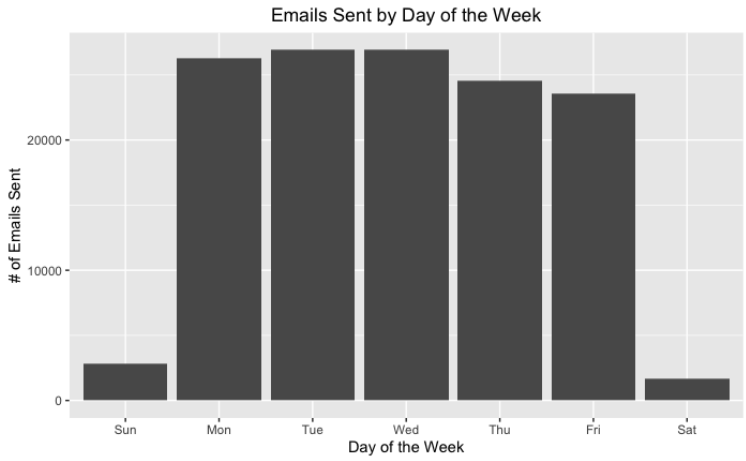
In Figure 3, note that the axes differ for each year, and that the bars are colored in accordance with business quarter. Very few emails are sent in 1999; this could be because only a subset of the relevant users were employed by Enron at that point, because few of those emails pertained to the scandal, etc. In 2002, few emails are sent after week 7, and by Q2 communication over these email addresses is effectively extinct.

Notably, toward the start of Q3 of 2001 there is a large spike in email communication. This is early October, when the Enron scandal was made public. There is also a suspicious drop-off in communication in June of 2001.

3.2.2 Emails sent by day of the week

When do employees send the most emails? Generally speaking, more emails should be sent on weekdays than on weekends. While the full dataset may be sloppier in terms of time zones (since only the 131 users' time zones are listed, not necessarily the senders'), the full dataset is more likely to represent overarching email habits. In contrast, the smaller dataset may be disproportionately bent toward the habits of the biggest sender of the group.

Figure 4: Emails sent by day of the week

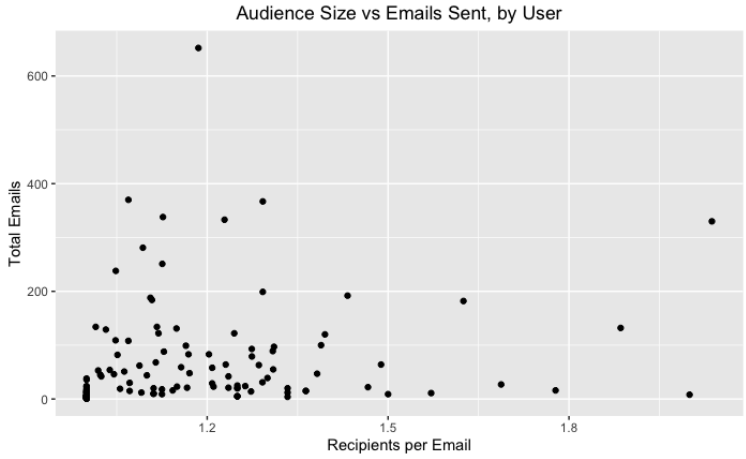


From Figure 4, it appears that Saturdays are the least common day for sending emails, and that Monday, Tuesday, and Wednesday are busier than Thursday and especially Friday.

3.3 Recipients per Email

Do the most prolific email senders have a larger or smaller email audience on average? Each dot is one of the users whose folders are in the dataset. Additionally, "recipients per email" considers only those users as recipients.

Figure 5: Users' audience size vs # of emails sent



Most users have relatively few emails and few recipients per email. There are extremes on both ends, but no correlation is particularly striking. The upper right corner of the chart is rather empty, which is not very surprising since too many people in a network who send lots of emails to lots of people will flood everyone's inboxes with emails.

3.4 Subject Lines Word Frequency

Which words are most commonly used in subject lines? The set of unique subject lines contains 444,399 words. The most common words are:

Rank	Word	Frequency	Rank	Word	Frequency	Rank	Word	Frequency
1	fw	10581	11	01	3217	21	report	1800
2	for	9180	12	2001	3056	22	from	1732
3	of	5401	13	in	2638	23	with	1542
4	to	5026	14	gas	2377	24	e	1524
5	and	4958	15	new	2349	25	call	1521
6	s	4046	16	1	2045	26	agreement	1516
7	the	3929	17	update	2022	27	2	1466
8	enron	3850	18	power	1916	28	request	1460
9	meeting	3828	19	energy	1874	29	11	1459
10	on	3570	20	a	1829	30	conference	1445

The most common word "fw" clearly comes from forwards, numbers come from dates, and both energy-related words and business-related words are prevalent.

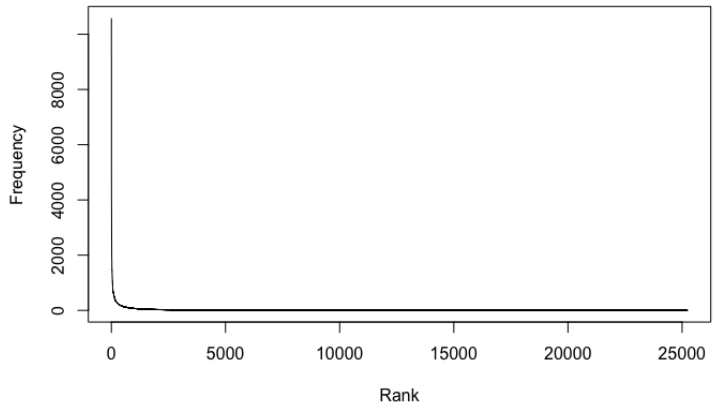
Amusingly, the six longest words are:

1. `bbbbbbbbbbbbbbbrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrr`
2. `chrissssssssssssssssssyyyyyyyyyyyyyyyyyyyyyy`
3. `chrissssssssssssssssssyyyyyyyyyyyyyyyyyyy`
4. `00000000000000000000000000000000000weil`
5. `petroleumworldweeklyreview`
6. `chrissssssssssyyyyyyyyyy`

since somebody evidently enjoyed bothering Chrissy.

Do subject lines abide by Zipf's Law? Clearly, people do not speak in the ways they tend to write their email subject lines. The subject lines are full of keywords, which disproportionately reflect words for key aspects of the business. I would expect words like "urgent" and "meeting" to be significantly over-represented in subject lines, compared with regular speech. I plotted frequency and rank.

Figure 6: Frequency and Rank of words in subject lines

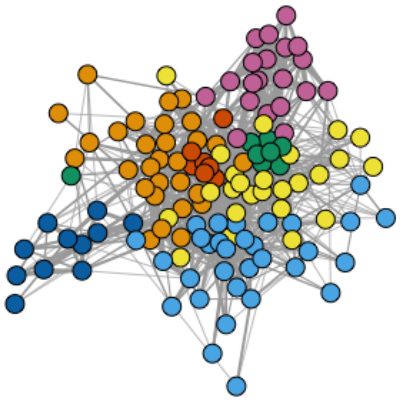


3.5 Networks

3.5.1 Graph

How interconnected are these 131 users? Can subgroups of users be clearly identified—for example, groups of people who collaborate frequently in their roles?

Figure 7: Network of correspondence between users



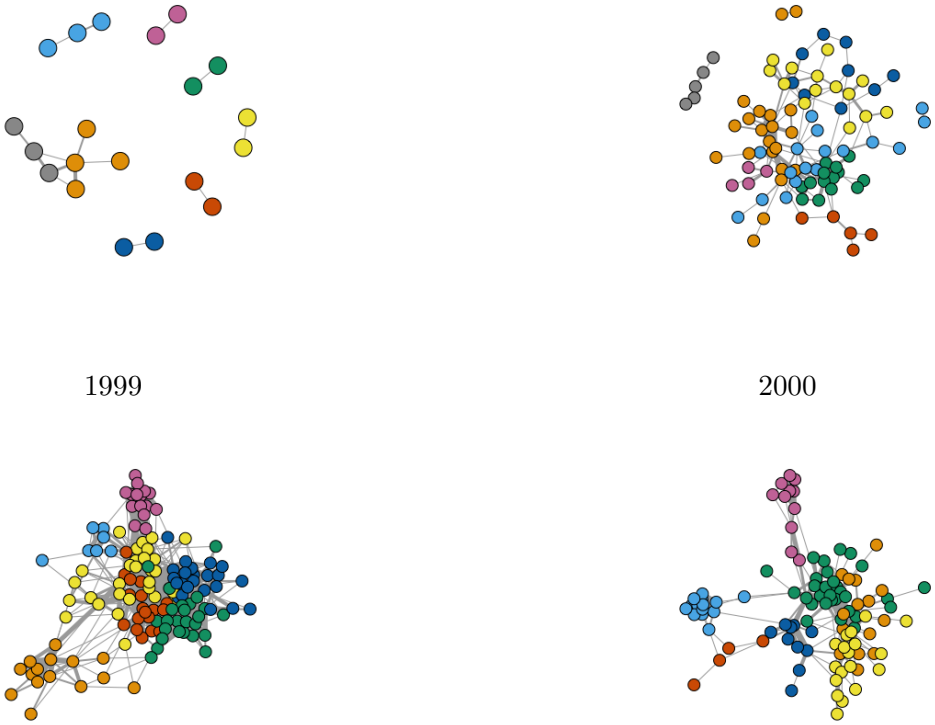
Each user is a vertex, and each graph edge is weighted by the natural log of number of correspondences between a pair of users. (I chose the log here because the otherwise the variation in

edge width was so extreme that it was difficult to see a complete picture.) Each vertex is colored by his or her community, which was assigned by iGraph's `cluster_fast_greedy()` function.

3.5.2 Networks over time

How did this network look in 1999, 2000, 2001, and 2002?

Figure 8: Network by Year, 1999-2002



1999 2000

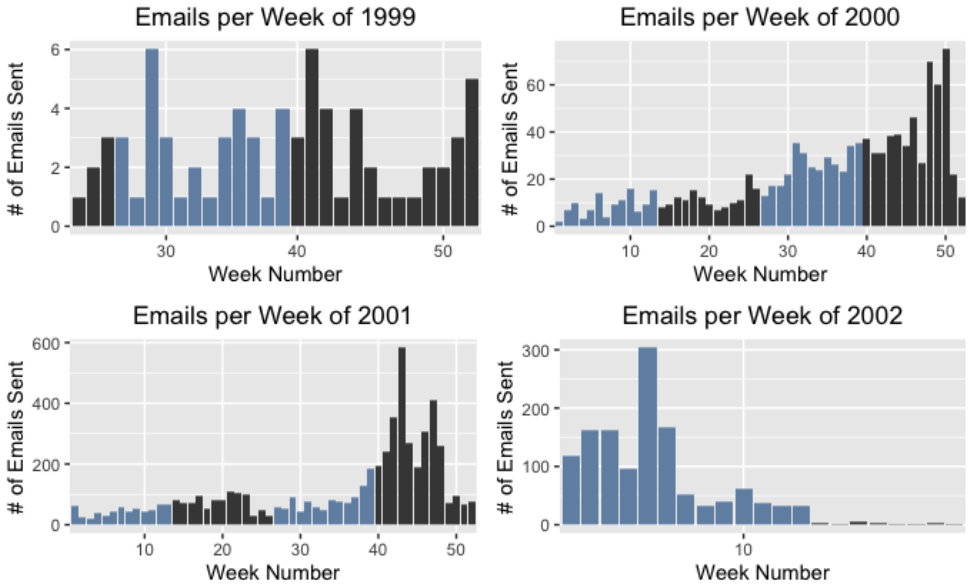
2001 2002

The graph in 1999 was sparse; only a handful of users emailed more than one person in this dataset. While 2000 had many more users and interactions than 1999, there were not easily identifiable clusters. Many users only sent a few emails to one or two other users, which do not result in clique-like structures. In 2001, users were much more heavily interconnected and formed clearer groups in some areas. By 2002, fewer emails were sent than in 2001 (and more than in 2000), but somewhat tightly knit groupings of users persisted.

These visual observations are substantiated by modularity calculations; the modularity was 0.543, 0.593, 0.552, and 0.623 in the years from 1999 to 2002, respectively. In comparison, the graph including all four years had a modularity score of 0.612. The bar plot for emails over time

(within the network dataset only) offers some context for the amount of correspondence users had in each year.

Figure 9: Emails sent by week of each year in the network dataset



Clearly, the bulk of the correspondence among users occurred in Q4 of 2001. To some degree, the 2001 network approximates the network during the scandal. Then, the change in the users' relationships from 2000 to 2002 can be partly explained by the chaos during the scandal.

In order to understand how and why the network shifted, I could analyze the content of emails over time and observe how frequently users corresponded over time. Did Q4 of 2001 encourage different employees to collaborate? When did particular employees quit, and did someone new fill vacated roles? These questions, and many others, could be explored given more time.