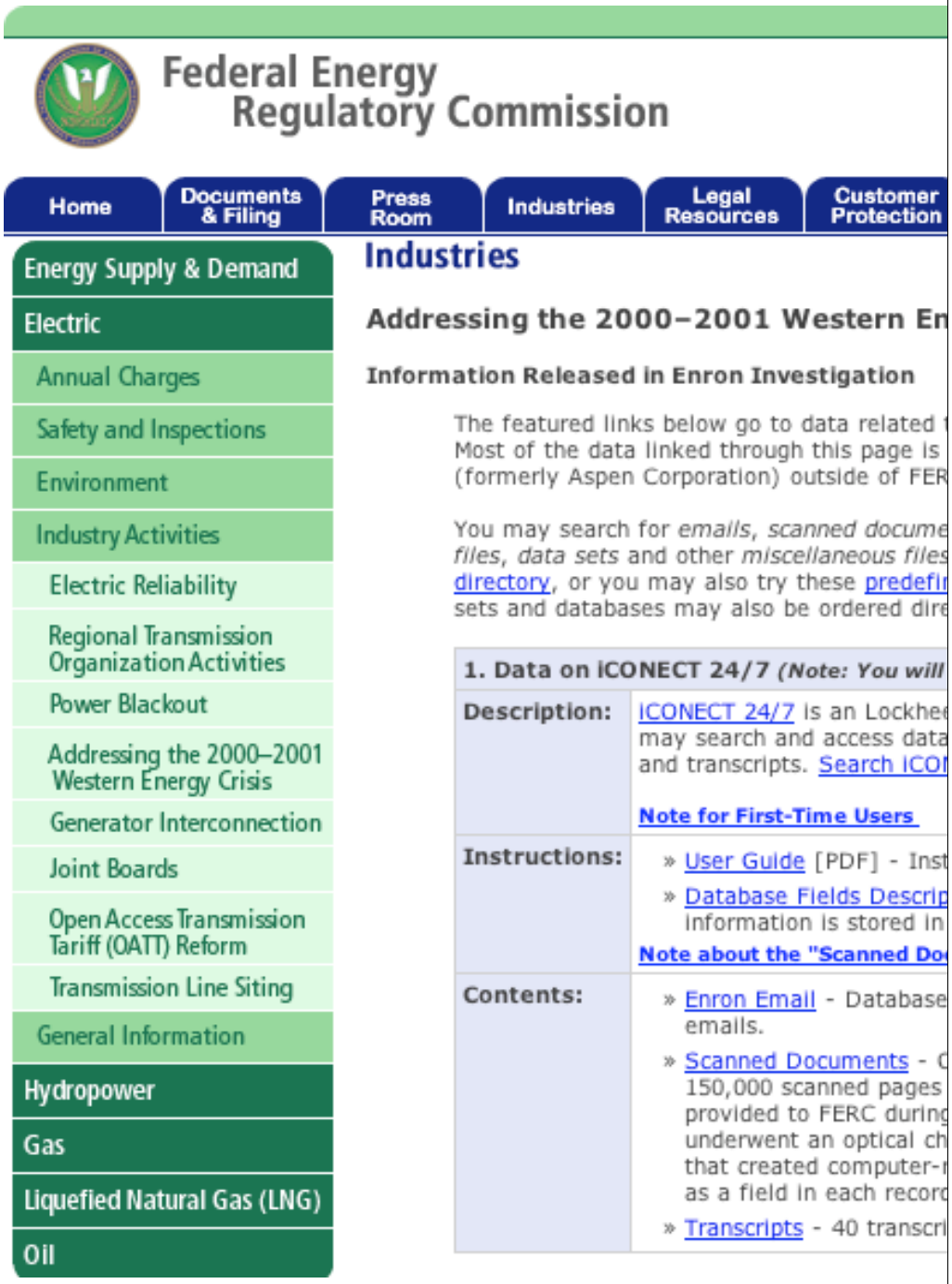
Example assignment:

Enron emails

As part of its investigation into Enron, the Federal Energy Regulatory Commission released the emails of about 150 of its top executives

These data were then cleaned up by groups at MIT and SRI and are now publicly available through the CMU CS Department

To respect the privacy of the individuals involved, I have replaced the body of each email with x's; our interest is not in what was said but who sent email to whom



Enron

Today we are going to start our work on a set of data related to the Enron corporation

Some relevant links are

http://www.chron.com/news/specials/enron/timeline.html

http://www.cs.cmu.edu/~enron/

http://www.stat.ucla.edu/~cocteau/klimt-ecml04-1.pdf

http://www.stat.ucla.edu/~cocteau/Enron_Employee_Status.htm

Organization of the data

The data itself is organized into a series of directories, each named after an executive

Under each directory, you will find possibly more directories, each representing a different mail folder

At the lowest level, you have a series of email messages, one per file; the files in each directory are named 1., 2., 3., etc.

xterm

```
[fad-gadget maildir] ls
allen-p          fischer-m        kitchen-l        phanis-s         smith-m
arnold-j         forney-j         kuykendall-t     pimenov-v        solberg-g
arora-h          fossum-d         lavorato-j       platter-p        south-s
badeer-r         gang-l           lay-k            presto-k         staab-t
bailey-s         gay-r            lenhart-m        quenet-j         stclair-c
bass-e           geaccone-t       lewis-a          quigley-d        steffes-j
baughman-d       germany-c        linder-e         rapp-b           stepenovitch-j
beck-s           gilbertsmith-d   lokay-m          reitmeyer-j      stokley-c
benson-r         giron-d          lokey-t          richey-c         storey-g
blair-l          griffith-j       love-p           ring-a           sturm-f
brawner-s        grigsby-m        lucci-p          ring-r           swerzbin-m
buy-r            guzman-m         maggi-m          rodrique-r       symes-k
campbell-l       haedicke-m       mann-k           rogers-b         taylor-m
carson-m         hain-m           martin-t         ruscitti-k       tholt-j
cash-m           harris-s         may-l            sager-e          thomas-p
causholli-m      hayslett-r       mccarty-d        saibi-e          townsend-j
corman-s         heard-m          mcconnell-m      salisbury-h      tycholiz-b
crandell-s       hendrickson-s    mckay-b          sanchez-m        ward-k
cuilla-m         hernandez-j      mckay-j          sanders-r        watson-k
dasovich-j       hodge-j          mclaughlin-e     scholtes-d       weldon-c
davis-d          holst-k          merriss-s        schoolcraft-d    whalley-g
dean-c           horton-s         meyers-a         schwieger-j      whalley-l
delainey-d       hyatt-k          mims-thurston-p  scott-s          white-s
derrick-j        hyvl-d           motley-m         semperger-c      whitt-m
dickson-s        jones-t          neal-s           shackleton-s     williams-j
donoho-l         kaminski-v       nemec-g          shankman-j       williams-w3
donohoe-t        kean-s           panus-s          shapiro-r        wolfe-j
dorland-c        keavey-p         parks-j          shively-h        ybarbo-p
ermis-f          keiser-k         pereira-s        skilling-j       zipper-a
farmer-d         king-j           perlingiere-d    slinger-r        zufferli-j
[fad-gadget maildir]
```

An example

Here we select the ex-Vice President for Regulatory Affairs, Shelley Corman

We see the 11 mail folders; selecting the calendar folder, we exhibit the content of mail 2.

Note again, that all textual content has been replaced by x's; we are only interested in (at best) the pattern of communication

xterm

```
[fad-gadget maildir] cd corman-s/
[fad-gadget corman-s] ls
1.                      contacts                ingaastudy
all_documents           deleted_items           marketingaffiliate
calendar                discussion_threads      osha
communications          inbox                   sent_items
[fad-gadget corman-s] cd calendar/
[fad-gadget calendar] ls
1.      19.     29.     38.     47.     56.     65.     74.     83.     92.
10.     2.      3.      39.     48.     57.     66.     75.     84.     93.
11.     20.     30.     4.      49.     58.     67.     76.     85.     94.
12.     21.     31.     40.     5.      59.     68.     77.     86.     95.
13.     22.     32.     41.     50.     6.      69.     78.     87.     96.
14.     23.     33.     42.     51.     60.     7.      79.     88.     97.
15.     25.     34.     43.     52.     61.     70.     8.      89.
16.     26.     35.     44.     53.     62.     71.     80.     9.
17.     27.     36.     45.     54.     63.     72.     81.     90.
18.     28.     37.     46.     55.     64.     73.     82.     91.
[fad-gadget calendar] cat 2.
Message-ID: <8257359.1075858837944.JavaMail.evans@thyme>
Date: Mon, 29 Oct 2001 10:23:04 -0800 (PST)
From: jean.mcfarland@enron.com
To: jean.mcfarland@enron.com, lynn.blair@enron.com, sheila.nacey@enron.com,
        john.buchanan@enron.com, toby.kuehl@enron.com,
        shelley.corman@enron.com, scott.abshire@enron.com,
        gary.kenagy@enron.com, bradley.holmes@enron.com, bob.hagen@enron.com,
        mary.vollmer@enron.com, terry.kowalke@enron.com,
        steve.january@enron.com, don.daze@enron.com
Subject: Updated: Overall Update for DRA (BCP)
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: McFarland, Jean </O=ENRON/OU=NA/CN=RECIPIENTS/CN=JMCFARL>
X-To: McFarland, Jean </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Jmcfarl>, Blair, Lynn <
=ENRON/OU=NA/CN=RECIPIENTS/CN=Lblair>, Nacey, Sheila </O=ENRON/OU=NA/CN=RECIPI
TS/CN=Snacey>, Buchanan, John </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Jbuchan2>, Kueh
 Toby </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Tkuehl>, Corman, Shelley </O=ENRON/OU=N
CN=RECIPIENTS/CN=Scorman>, Abshire, Scott </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Sab
ir>, Kenagy, Gary </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Gkenagy>, Holmes, Bradley <
=ENRON/OU=NA/CN=RECIPIENTS/CN=Bholmes>, Hagen, Bob </O=ENRON/OU=NA/CN=RECIPIEN
/CN=Bhagen>, Vollmer, Mary </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Mvollme>, Kowalke,
erry </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Tkowalk>, January, Steve </O=ENRON/OU=NA
N=RECIPIENTS/CN=Sjanuary>, Daze, Don </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Ddaze>
X-cc:
X-bcc:
X-Folder: \SCORMAN (Non-Privileged)\Calendar
X-Origin: Corman-S
X-FileName: SCORMAN (Non-Privileged).pst

xx xxxxx xxxx xx xxxx xxxxxxx xx xxxxxx xx xxxxxxxxxx xxxxxxxxxxx xxx xxx (xx
.   xxxxxx xxxx xxxx xx xxxxxx xxxxxx xx xxx xxxx xxxxxx xxxx xxx xxxxxxxxxxxx
xxx.

xxxxxx.  xxxx xxxxxx
[fad-gadget calendar]
```

Some questions

What is the distribution of numbers of emails per user?

Are the users organizing their email into folders?

Are certain folders common to all users?

What is the distribution of emails per folder?

Some questions

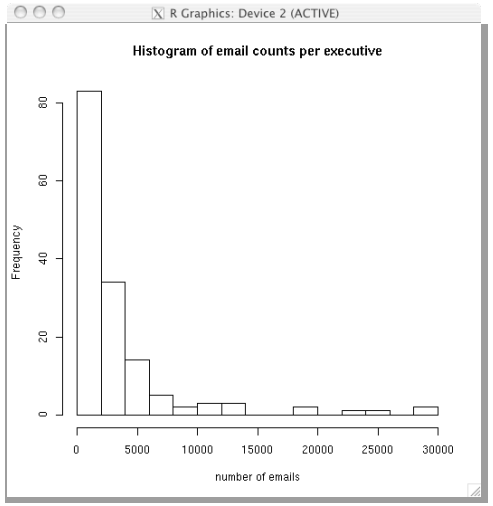
These first questions can be addressed mainly through the use of Unix shell commands (`cut`, `sort`, `uniq`, `grep`) for a single user

To iterate over all of the folders, we introduce `find` and simple shell scripting; this forces us to look at the file system a little more closely

Given data in this (admittedly) horrible format, we find motivation for a number of lectures on the basics of Unix

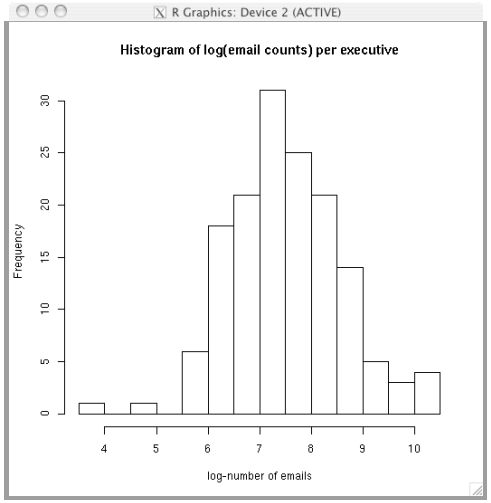
Counts per user

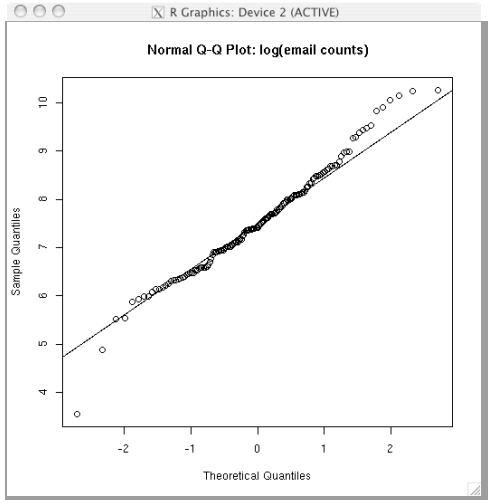
As was the case for hit counts per IP address, we see a very skewed distribution (what Malcolm Gladwell would call a “hockey stick” distribution)



In the bottom figures we present a histogram and a Q-Q plot for the logarithm of the counts

We show these plots now but technically, the students encounter R after this initial foray





Abstractions

In my course, Unix leads to a general purpose scripting language (at first it was Perl, now Python)

The Enron example can easily follow along, providing students with their first glimpse of how data might be organized and accessed in a more formal way

A taste

Suppose we want to extract the `To:` and `From:` information from each of the 0.5M emails

xterm

```
Message-ID: <548548.1075861083665.JavaMail.evans@thyme>
Date: Fri, 22 Mar 2002 11:59:25 -0800 (PST)
From: dale.m.davis@williams.com
To: griffith'.'bill@enron.com, fava'.'gene@enron.com, king'.'iris@enron.com,
        keeler'.'john@enron.com, burch'.'kathryn@enron.com,
        pelt'.'kim@enron.com, schubert'.'ken@enron.com,
        mccain'.'marcy@enron.com, gracey'.'mark@enron.com,
        wilke'.'mark@enron.com, love'.'paul@enron.com,
        young'.'randy@enron.com, theresa.hess@enron.com,
        gwilliam'.'tom@enron.com, grygar'.'bill@enron.com,
        charlie.bass@enron.com, shelley.corman@enron.com
Subject: Order RM96-1-019 (Partial Day Recalls) - BPS Schedule is Official &
 Pipeline Segment Meeting
Cc: 8772405711@pagenetmessage.net
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: 8772405711@pagenetmessage.net
X-From: Davis, Dale M <Dale.M.Davis@Williams.com>
X-To: 'Bill Griffith' <william.griffith@elpaso.com>, 'Gene Fava' <efava@glgt.com
>, 'Iris King' <iris_g._king@dom.com>, 'John Keeler' <jkeeler@glgt.com>, 'Kathry
n Burch' <klburch@duke-energy.com>, 'Kim Van Pelt' <kvanpelt@cmsenergy.com>, 'Ke
n Schubert' <ken_schubert@transcanada.com>, 'Marcy McCain' <mlmccain@duke-energy
[id-55-241:maildir/corman-s/inbox] cocteau%
```

A taste

A few lines of Python will do the trick; in short, we create a object that encapsulates the characteristics of an email message and then work with that object

```
xterm
[id-55-241:maildir/corman-s/inbox] cocteau% pwd
/Users/Shared/data/maildir/corman-s/inbox
[id-55-241:maildir/corman-s/inbox] cocteau%
[id-55-241:maildir/corman-s/inbox] cocteau% python
Python 2.3.5 (#1, Mar 20 2005, 20:38:20)
[GCC 3.3 20030304 (Apple Computer, Inc. build 1809)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
>>> import email
>>>
>>> msgfile = open("53.")      # open the 53rd email message file
>>> msg = email.message_from_file(msgfile)
>>> msgfile.close()
>>>
>>> msg['Subject']
'Thank s'
>>> msg['To']
'shelley.corman@enron.com'
>>> msg['From']
'stanley.horton@enron.com'
>>>
```

The final assignment

Ultimately (after we get to R), students are asked to assess whether the executives' communication patterns change over time; there are several ways to do this, some of which involve metrics derived from social networks

I am not overly picky about what the students choose to compute from these data, but I am fussy about process; I want them to accompany their "answer" with some sense of how they know they're right

By that I mean, not only produce a number or a graph, but also illustrate how the information they extracted from the directory tree is what they expect it to be