## SIMS 290 – Applied NLP  Assignment 4: Enron Email Corpus

Eva Mok (emok@icsi.berkeley.edu)

**Project goal:**

Preliminary social network analysis on the Enron data, based on the number of emails sent between people in the organization (i.e. no information extracted from email body).

**Methodology:**

To limit the size and scope of this project, the dataset I used contains only the subset of the Enron email corpus that have been annotated by the class. The annotation, however, is not used – the information extracted from each email comes strictly from the header section of the email. Two kinds of information are extracted: (1) sender and recipient email addresses for each email (used for email counts), and (2) email addresses and the corresponding name, if available. For this purpose, both the header and X-header sections are used. The result of the processing is a weighted directed graph (represented as an edgelist) of email counts between sender and recipient pairs, e.g. an email from one person to 5 recipients results in a graph of 6 nodes and 5 edges, each with weight 1. The individuals are identified by his/her email address[1]. In addition, a dictionary of email-name pairs is constructed. This dictionary is not used directly for the social network analysis but is used to help understand the results.

For the social network analysis, I used UCINET 6, a software developed by Borgatti, Everett, and Freeman, and distributed by Analytic Technologies. The choice of the software was somewhat arbitrary, and should be considered an exploration step – I was mainly looking for a package that would both visualize data as well as perform various analysis. The analysis that I have eventually performed on the network include: cliques, N-cliques, degree centrality, closeness centrality, betweenness centrality, ego network, and core/peripheral groups. I will discuss these in more details in a later section.

*Data processing phase:*

The header information is obtained using enronEmail.py written by Andrew T. Fiore. To get the email counts and names, eight fields in the header are used: *To, From, Cc, Bcc, X-To, X-From, X-Cc, X-Bc*c. For the purpose of this project, recipients in the *To, Cc* and *Bcc* fields are treated equally, and redundancy in the fields is not eliminated (e.g. someone who is both in the *To* list and the *Cc* list will be counted twice). This information is kept in a conditional frequency distribution.

To get the email-name pairs, I tried to match the information in the X-header to the header. The X-header information comes in various formats, and is not always present (or complete). The ordering of the X-header entries, I observed, seems to generally correspond to the ordering in the basic header. Eventually I used a few regular expressions to get rid of unnecessary information and to split up the lines into names, make sure that the number of entries in the X-header matches that of the basic header, and paired up the emails and names.

[1] Apparently this assumption falls apart, since Vince J Kaminski apparently has a number of email aliases and addresses, such as vince.j.kaminski@enron.com, vince.kaminski@enron.com, j.kaminski@enron.com, vkaminski@aol.com, kaminski@enron.com

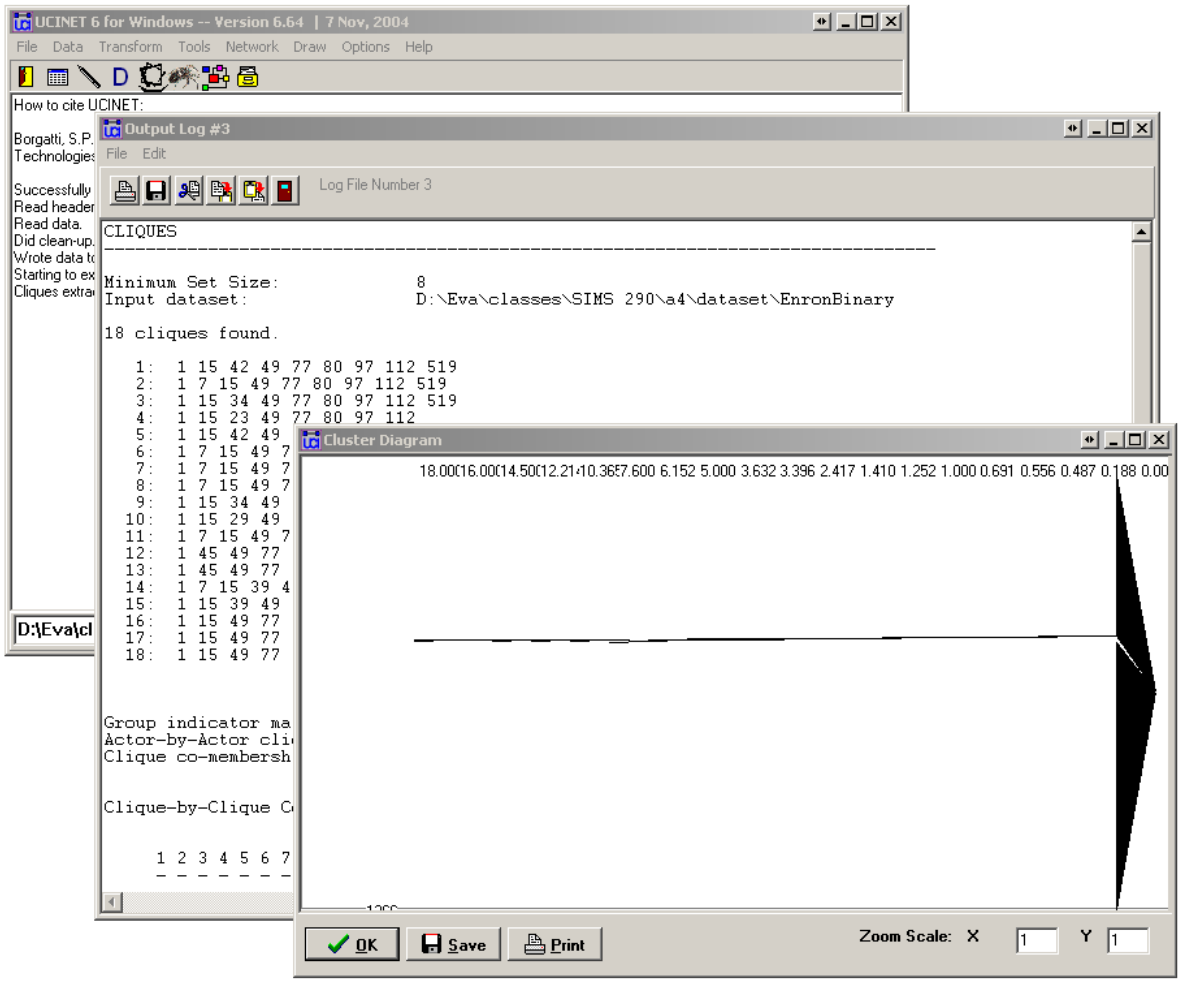
*Social network analysis:*

UCINET can import data in a number of formats, and the edgelist format is used in this project. This format looks like:

```
dl n=1366 format=edgelist1
data:
1 2 2
1 3 1
1 4 3
1 5 1
…
```

where n is the number of nodes in the network, and each line in the data section represents the two connecting nodes and the weight on that edge.

Certain analysis requires symmetric matrix, and certain other analysis requires a binary matrix, so the data is further processed once loaded into UCINET to obtain (1) a symmetric version (by summing the number of email counts of the upper and lower matrices, reflecting the total number of correspondence between two people), and (2) a binary version (by forcing the symmetric version to 0 and 1, reflecting the presence of communication between two people).



**Results:**

The graph I obtained from this subset of the email corpus is a sparsely connected directed graph of 1366 nodes and 2454 edges. The graph contains unconnected subgraphs (according to the closeness centrality computation).

*cliques:*

Based on the binary version of the graph, I looked for cliques. There are 70 cliques of size 7 or larger, and 18 cliques of size 8 or larger. The largest clique is size 9 (5 of them). Here are the cliques of minimum size 8:

```
Minimum Set Size:     8
Input dataset:        D:\Eva\classes\SIMS 290\a4\dataset\EnronBinary

18 cliques found.

   1:  1 15 42 49 77 80 97 112 519
   2:  1 7 15 49 77 80 97 112 519
   3:  1 15 34 49 77 80 97 112 519
   4:  1 15 23 49 77 80 97 112
   5:  1 15 42 49 77 80 97 489
   6:  1 7 15 49 77 80 97 489
   7:  1 7 15 49 77 80 447 519 524
   8:  1 7 15 49 77 80 447 489
   9:  1 15 34 49 77 80 447 519
  10:  1 15 29 49 80 447 519 524
  11:  1 7 15 49 77 80 112 519 524
  12:  1 45 49 77 80 112 519 524
  13:  1 45 49 77 80 447 519 524
  14:  1 7 15 39 49 447 519 524
  15:  1 15 39 49 89 447 519 524
  16:  1 15 49 77 89 97 112 519
  17:  1 15 49 77 89 112 519 524
  18:  1 15 49 77 89 447 519 524
```

Here is a sample of the people in the cliques:

1: Susan J Mara, Alan Comnes, Sandra McCubbin, Steven J Kean, James D Steffes, Jeff Dasovich, Mary Hain, Karen Denne, Miyung Buster

5: Susan J Mara, Alan Comnes, Sandra McCubbin, Steven J Kean, James D Steffes, Jeff Dasovich, Mary Hain, Tom Hoatson

12: Susan J Mara, Janel Guerrero, Steven J Kean, James D Steffes, Jeff Dasovich, Karen Denne, Miyung Buster, Angela Wilson

*N-cliques:*

In addition to directly connected cliques, I also looked for 2-cliques, and the results that come out indicates that there are 437 cliques of size 8 or larger, with the biggest clique being 358. About 400 of these are under size of 100.

*Degree Centrality:*

Using the original (directed, weighted) graph, I computed the degree centrality of the network, which reflects the amount of ties an individual has within the network. Here is a partial list of the results I obtained:

```
Diagonal valid?               NO
Model:                        ASYMMETRIC
Input dataset:                D:\Eva\classes\SIMS 290\a4\dataset\Enron
```

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | OutDegree | InDegree | NrmOutDeg | NrmInDeg |
| 49 | 2252.000 | 83.000 | 164.982 | 6.081 |
| 519 | 1534.000 | 8.000 | 112.381 | 0.586 |
| 447 | 907.000 | 20.000 | 66.447 | 1.465 |
| 1042 | 886.000 | 2.000 | 64.908 | 0.147 |
| 15 | 430.000 | 64.000 | 31.502 | 4.689 |
| 437 | 268.000 | 4.000 | 19.634 | 0.293 |
| 1 | 203.000 | 116.000 | 14.872 | 8.498 |
| 80 | 187.000 | 208.000 | 13.700 | 15.238 |
| 97 | 181.000 | 38.000 | 13.260 | 2.784 |
| 567 | 179.000 | 0.000 | 13.114 | 0.000 |

The top ten people shown in the above list correspond to:
Steven J Kean, Miyung Buster, John Shelk, Jean Munoz (jmunoz@mcnallytemple.com), Alan Comnes, Vince J Kaminski, Susan J Mara, Jeff Dasovich, Mary Hain, bwoertz@caiso.com.

*Closeness Centrality:*

Using the binary graph, I computed the (normalized) closeness centrality. The closeness centrality is defined as the reciprocal of farness, which is the sum of the lengths of the geodesics to every other vertex. Here is a partial list of the results I obtained:

```
Input dataset:        D:\Eva\classes\SIMS 290\a4\dataset\EnronBinary
Method:               Geodesic paths only (Freeman Closeness)
Output dataset:       D:\Eva\classes\SIMS 290\a4\dataset\Closeness

The network is not connected. Technically, closeness centrality
cannot be computed, as there are infinite distances.


Closeness Centrality Measures
```

| | 1 | 2 |
|---|---|---|
| | Farness | nCloseness |
| 49 | 44157.000 | 3.091 |
| 80 | 44538.000 | 3.065 |
| 1 | 44648.000 | 3.057 |
| 15 | 44681.000 | 3.055 |
| 97 | 44797.000 | 3.047 |
| 112 | 44833.000 | 3.045 |
| 519 | 44849.000 | 3.044 |
| 447 | 44864.000 | 3.043 |

```
 56     44870.000        3.042
524     44870.000        3.042
…

146   1863225.000        0.073
1364   1863225.000        0.073
1363   1863225.000        0.073
418   1863225.000        0.073
1164   1863225.000        0.073
```

The top 5 people on the closeness centrality measure are:
Steven J Kean, Jeff Dasovich, Susan J Mara, Alan Comnes, Mary Hain

And the least closely connected 5 are (in the order shown above):
Andrew H Lewis, IOS_Participants@ypo.org, Brenda_Worley@ypo.org, Caleb Offley (offley@hoover.stanford.edu), Bill Williams III

*Betweenness Centrality:*

Using the original directed graph, I looked at the betweenness centrality in the network, which reflects the value of an individual as an intermediary between other individuals. Here is the top 10 results I obtained:

```
Input dataset:        D:\Eva\classes\SIMS 290\a4\dataset\Enron

Important note: this routine binarizes but does NOT symmetrize.

Un-normalized centralization: 35506655.831

                1            2
     Betweenness nBetweenness
     ------------ ------------
 49     26106.748        1.402
 56     13218.109        0.710
 80     11743.800        0.631
 15     10944.622        0.588
 97      7517.736        0.404
  1      7257.391        0.390
 21      6375.403        0.342
179      4908.333        0.264
437      4731.667        0.254
581      4519.667        0.243
```

These people correspond to:
Steven J Kean, gfergus@brobeck.com, Jeff Dasovich, Alan Comnes, Mary Hain, Susan J Mara, Richard B Sanders, Vince J Kaminski (vince.kaminski@enron.com), Vince J Kaminski (j.kaminski@enron.com), Frank A. Wolak (wolak@zia.stanford.edu)

*Ego Network:*

With the original graph, I computed ego network density for each of the individual. A number of measures are computed for each individual, including size (the number of individuals that he is directly connected to), ties, pairs, density, etc. I ranked the individuals based on the size of his network (I did this in Excel), and here are the results I obtained:

| | Size | Ties | Pairs | Density | nWeakComp | pWeakComp | 2StepReach | ReachEffic | Broker | nBroker | EgoBetween | nEgoBetween |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | 353 | 489 | 124256 | 0.4 | 178 | 50.4 | 51.0 | 36.8 | 61883.5 | 0.5 | 8943.6 | 7.2 |
| 80 | 130 | 197 | 16770 | 1.2 | 36 | 27.7 | 24.4 | 36.2 | 8286.5 | 0.5 | 1160.9 | 6.9 |
| 15 | 123 | 350 | 15006 | 2.3 | 11 | 8.9 | 43.9 | 39.6 | 7328.0 | 0.5 | 1390.6 | 9.3 |
| 1 | 116 | 321 | 13340 | 2.4 | 7 | 6.0 | 43.7 | 40.9 | 6509.5 | 0.5 | 664.4 | 5.0 |
| 1206 | 103 | 1 | 10506 | 0.0 | 102 | 99.0 | 9.3 | 81.4 | 5252.5 | 0.5 | 0.0 | 0.0 |
| 97 | 86 | 271 | 7310 | 3.7 | 3 | 3.5 | 43.3 | 43.0 | 3519.5 | 0.5 | 434.3 | 5.9 |
| 447 | 84 | 178 | 6972 | 2.6 | 32 | 38.1 | 44.2 | 49.1 | 3397.0 | 0.5 | 652.4 | 9.4 |
| 519 | 82 | 256 | 6642 | 3.9 | 10 | 12.2 | 45.0 | 42.7 | 3193.0 | 0.5 | 589.1 | 8.9 |
| 1042 | 82 | 47 | 6642 | 0.7 | 36 | 43.9 | 17.4 | 65.4 | 3297.5 | 0.5 | 35.0 | 0.5 |
| 437 | 72 | 6 | 5112 | 0.1 | 66 | 91.7 | 6.1 | 83.8 | 2553.0 | 0.5 | 209.0 | 4.1 |

And these people are:
Steven J Kean, Jeff Dasovich, Alan Comnes, Susan J Mara, dan.wall@lw.com, Mary Hain, John Shelk, Miyung Buster, Jean Munoz (jmunoz@mcnallytemple.com), Vince J Kaminski.

*Core/Periphery:*

Finally, I used the original graph to compute the core / periphery group, which clusters individuals into the two categories (this is done using a genetic algorithm in UCINET). This analysis identified 5 individuals in the core, and they are:

Richard Shapiro, Linda Robertson, Steven J Kean, James D Steffes, Jeff Dasovich

**Discussion:**

This project is a first step in performing social network analysis on the Enron organization based on the email corpus. As evident by the results, the analysis is heavily biased by the sample of emails used in the analysis. Further improvements would include a more robust way of identifying individuals in the organization from the emails (e.g. using a name entity recognizer), and/or exploiting information in the body of the emails (e.g. name mentions, or accounting for which emails are replied to).