Networks

MSPA PREDICT 455-DL-SEC55

Darryl Buswell

1 Introduction

This assignment explores text from Enron email corpus, with the aim to identify trends or anaomalies in employee communication or behaviour using network visualization techniques.

2 Data

The 'original' Enron email corpus was uploaded by William Cohen from CMU in March 2004 (Cohen 2015). This version of the corpus contains 517,431 distinct email from 151 users. The corpus does however have a number of blank and duplicate emails as well as junk system data from email transaction failures. Jitesh Shetty and Jafar Adibi from ISI later uploaded a MySQL4 version of the corpus which attempted to fix those problems. This version of the corpus retains only selected tables with duplicate emails removed and names normalized, resulting in 252,759 emails from 151 users. For this exercise however, I have elected to use a version of the corpus provided by Schulz, which is essentially a MySQL5 version of the dataset provided by Shetty and Adibi, with some additional data cleaning (Schulz 2015).

3 Data Exploration

We are able to query the MySQL5 dataset for both a table of nodes and edges. In this case, the nodes table contains information relevant to specific Enron employees, including email address, last name and employment status. A summary of the nodes table can be found in Appendix A. The edges table on the other hand, contains records for email communication between Enron employees, including the address the email was sent from, address the email was sent to, whether the email was directly sent, or whether the recipient was CC'd or BCC'd, and the date at which the email transaction occured. A summary of the edges table can also be found in Appendix A.

Using the table of nodes and edges, we are able to create an igraph network which contains the (symbolic) edge list and edge/vertex attributes. Using this network, we can create a histogram and cumulative plot of the degree of each node. In this case, the degree represents the number of emails sent to/from a particular employee. Figure B1 in Appendix B shows a histogram of node degrees, whilst Figure B2 shows the cumulative degree of each node. We can see that the majority of nodes carry a degree of 1000 or less.

We can also use the same igraph package to create a network plot of each node and their corresponding edges. We start by greating a network of the Enron email dataset, without any filters or graphical overlays. Figure B3 shows this unfiltered network. Clearly the lack of categorization and limited use of graphical overlays makes it difficult to make any inferences from this plot. Also note that this plot includes three nodes which do not have a corresponding edge.

In order to improve the network plot, we add a filter to remove nodes which have less than one degree, add a color overlay for each node to represent the position of the employee, size each of the nodes according to its degree (greater node sizes represent a greater degree), and finally, include only emails which were sent directly to a recipient, rather than all emails which were 'CC'd' or 'BCC'd'. Figure B4 shows this filtered network. Again, it is difficult to make any inferences from this network plot. However, it does seem that those with the status 'Vice President' tend to have larger node sizes (and therefore have a greater degree).

To further simplify the network plot, we can be more aggressive with the degree filter and exlude those nodes which have a degree less than 200. We also color each edge according to its origin node. Figure B5 shows this filtered network. Again, it is difficult to make any inferences from this network plot, however it is slightly more obvious that those with the status 'Vice President' tend to have larger node sizes, and that nodes with this status tend to email others with the same status.

In order to get an idea of how email communication at Enron changed over time, we can create a separate network graph for emails over 2000, 2001 and 2002. Do note that the dataset has significantly more emails sent over the year 2001. We retain the previously discussed filters for each of these network graphs. For each year, we also create a circle network plot in order to allow us to gain a better understanding of the communication links between each node type. These plots are shown in Figure B6 through to Figure B11. Unfortunately the

spread of data makes it difficult to draw comparisons between calendar years. An interesting extension to this work may be to subset emails sent over the year 2001, by quater, and to color only edges for one or two of the node types at a time. This may provide greater insights into any observable shifts in communication patterns for each node type over time.

To get a better idea of which individual nodes are most influential, we can apply hub and authority scores as developed by Jon Kleinberg (Kleinberg 1999). Nodes with a high hub score are expected to have a large number of outgoing emails while nodes with a high authority score are expected to have a large number of incoming emails. Figure B13 shows the unfiltered network plot with nodes sized according to their hub score. We can see a few nodes with a relatively large hub score and Table A3 shows the nodes with the top five hub scores. Figure B14 shows the unfiltered network plot with nodes sized according to their authority score. We also see a similar amount of nodes with a relatively high authority score and Table A4 shows the nodes with the fop five authority scores.

4 Conclusion

We were able to process the dataset in order to derive a network plot of emails at Enron. Although we applied a number of filtering techniques, we found it difficult to extract any meaningful insights from the plots themselves. It may be that these plots could be further improved on through alternative categorizations or by making alternative subsets. However at least for this assessment, the greatest insights were able to be made by simply applying a hub/authority metric in order to find those nodes with the greatest influence.

Appendix A Table Output

Table A1: Nodes Table Summary

Column Name	Type	Values
Email_id	chr	'albert.meyers@enron.com' ...
lastName	chr	'Taylor' 'Donoho' 'Gang' ...
status	chr	'N/A' 'Employee' ...

Table A2: Edges Table Summary

Column Name	Type	Values
sender	Factor w/ 144 levels	'albert.meyers@enron.com' ...
reciever	Factor w/ 146 levels	'albert.meyers@enron.com' ...
type	Factor w/ 3 levels	'BCC','CC','TO'
date	Date	'2002-01-25' '2002-01-24' ...

Table A3: Top Five Nodes Sorted by Hub Score

Node	Hubscore
jeff.dasovich@enron.com	1.00000000
james.d.steffes@enron.com	0.26304852
steven.j.kean@enron.com	0.13700000
richard.b.sanders@enron.com	0.10866270
mary.hain@enron.com	0.09921621
louise.kitchen@enron.com	0.02374020

Table A4: Top Five Nodes Sorted by Authority Score

Node	Authscore
richard.shapiro@enron.com	1.0000000
james.d.steffes@enron.com	0.9994337
steven.j.kean@enron.com	0.7529818
richard.b.sanders@enron.com	0.3811574
mary.hain@enron.com	0.1356280
robert.badeer@enron.com	0.1326668

Appendix B Figure Output

Figure B1 Degree Frequency

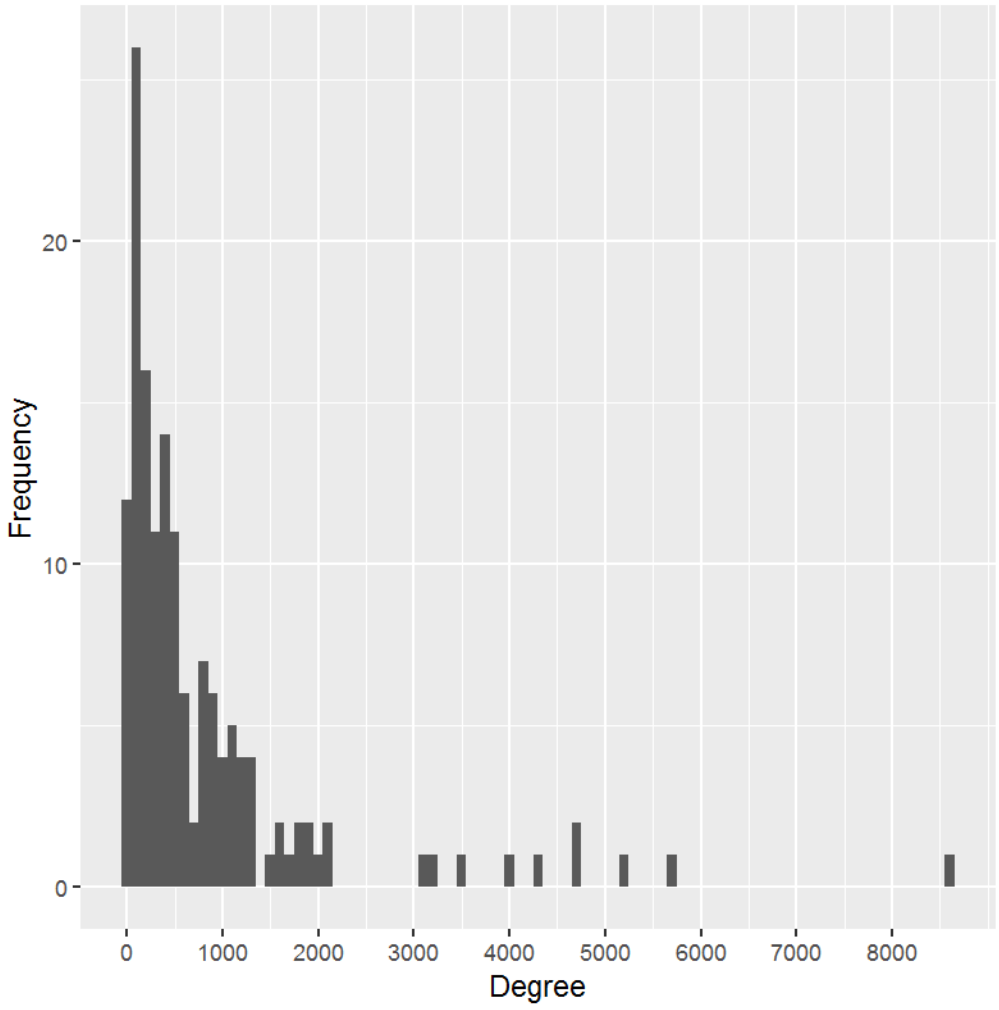


Figure B2 Degree Cumulative Frequency

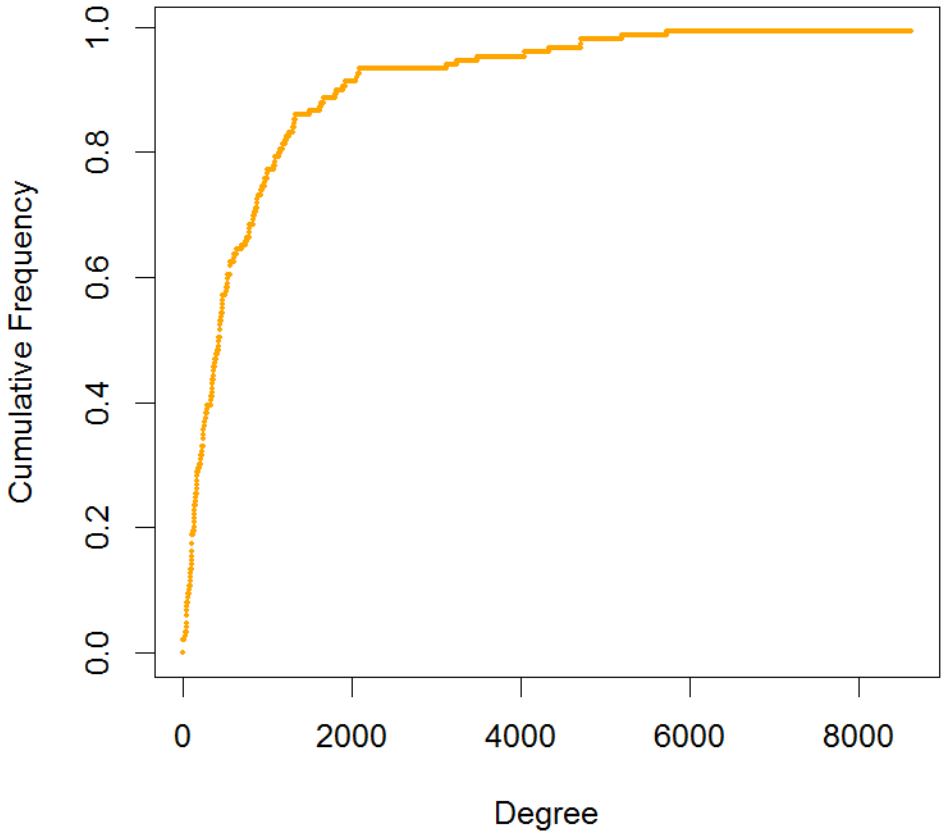


Figure B3 Raw Network of Enron Emails

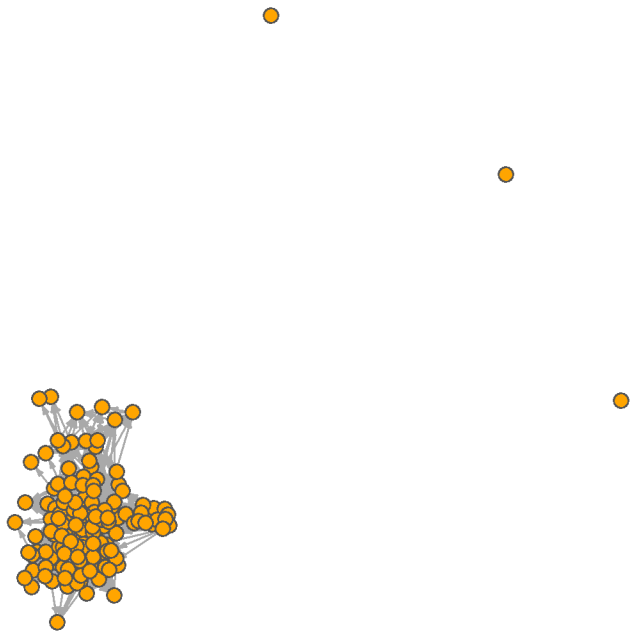


Figure B4 Network of Enron Emails (Emails sent TO)

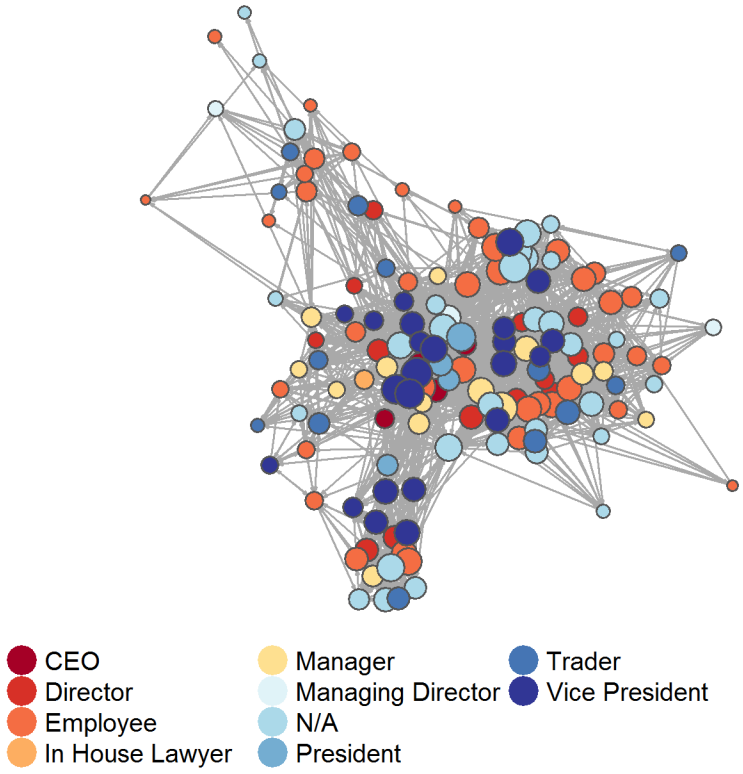


Figure B5 Network of Enron Emails (Greater than 200 Emails sent TO)

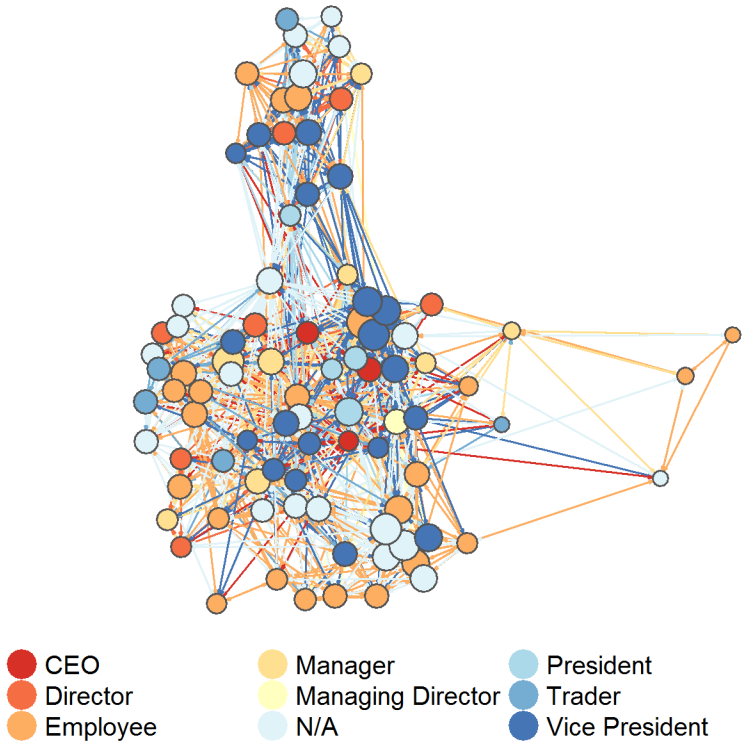


Figure B6 Network of Enron Emails (Emails sent TO over 2000)

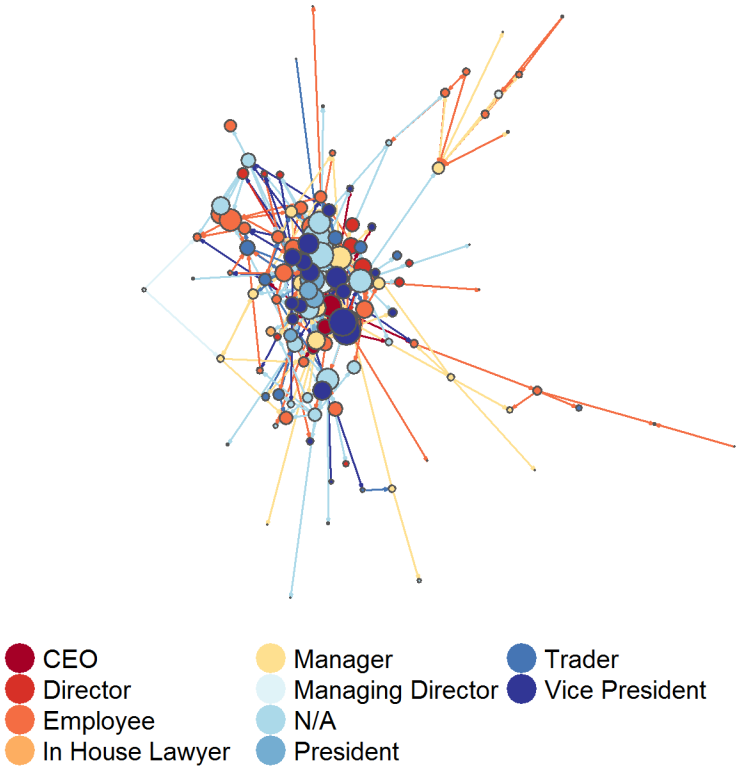


Figure B7 Network of Enron Emails (Emails sent TO over 2000)

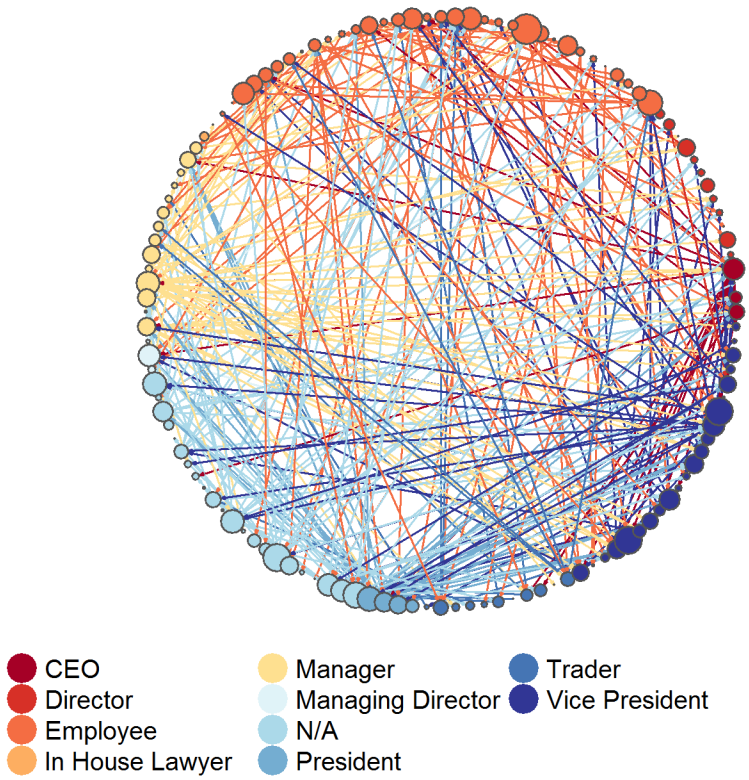


Figure B8 Network of Enron Emails (Emails sent TO over 2001)

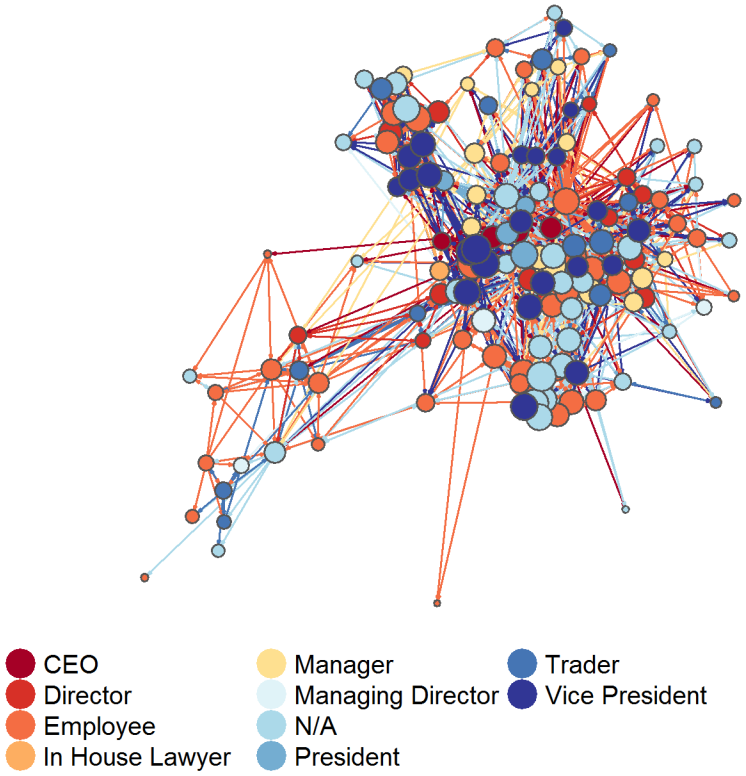


Figure B9 Network of Enron Emails (Emails sent TO over 2001)

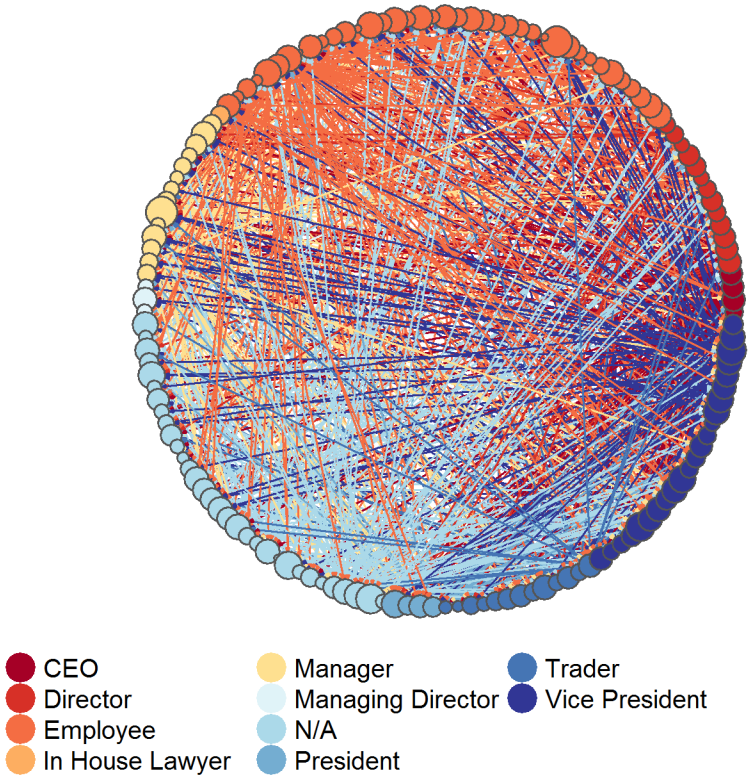


Figure B10 Network of Enron Emails (Emails sent TO over 2002)

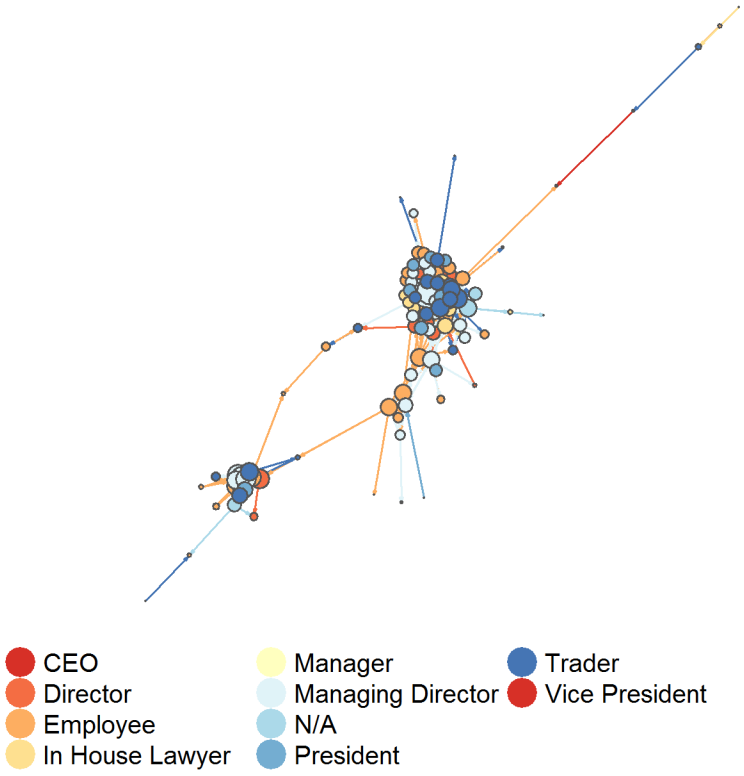


Figure B11 Network of Enron Emails (Emails sent TO over 2002)

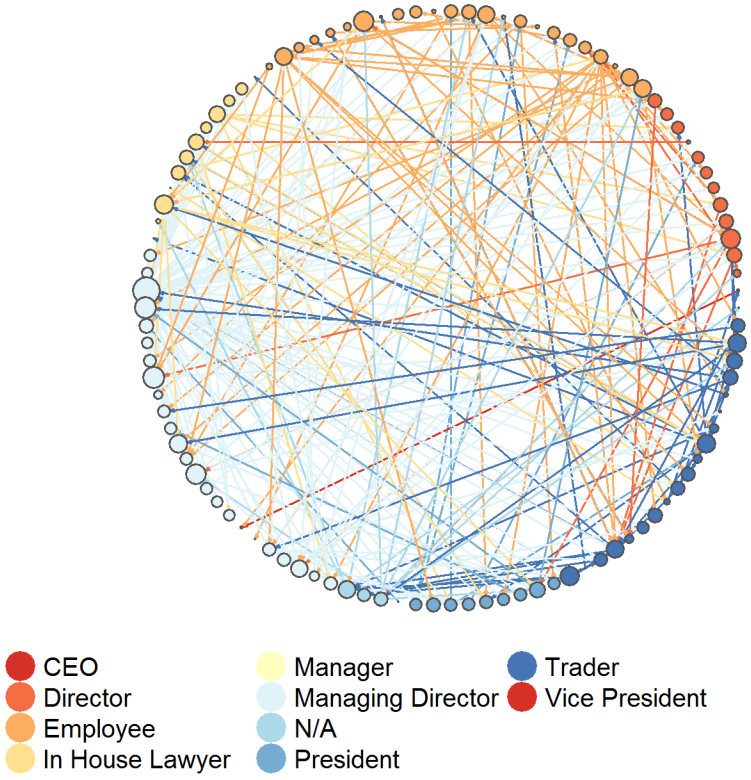


Figure B12 Network of Enron Emails (Nodes sized by Hub Score)

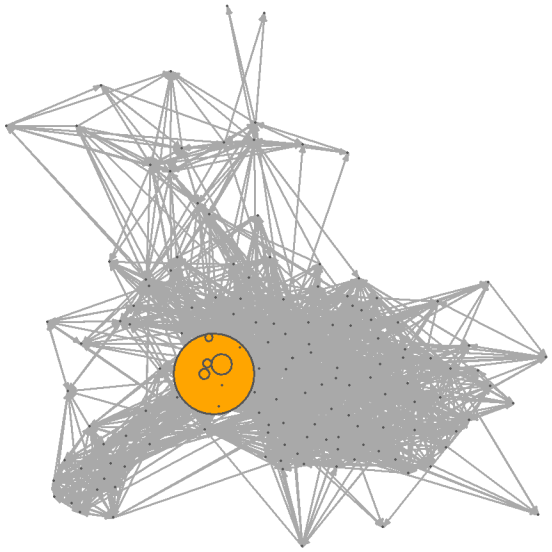
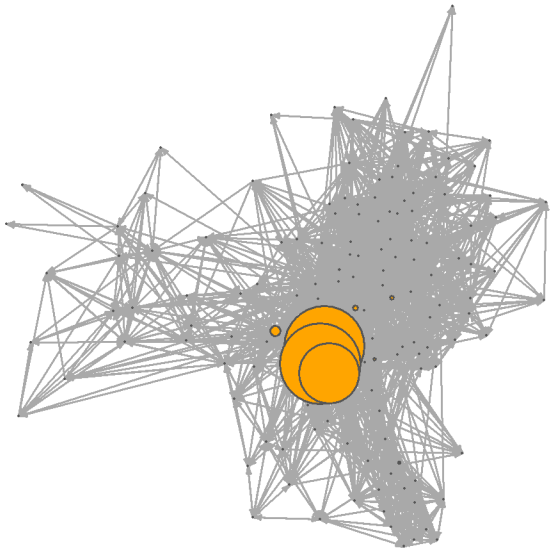


Figure B13 Network of Enron Emails (Nodes sized by Authority Score)



References

Cohen, W. 2015. “Enron Email Dataset.” http://www.cs.cmu.edu/~enron/.

Kleinberg, J. 1999. “Hubs, Authorities, and Communities.” http://cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html.

Schulz, A. 2015. “Enron Data.” http://www.ahschulz.de/enron-email-data/.